Estimating PATE under positivity violations: SBART+SPL for high-dimensional covariates

Lennard Maßmann*†

May 27, 2025

Abstract

The positivity assumption is a fundamental requirement for causal inference in the potential outcomes framework, ensuring that all individuals have a positive probability of receiving each treatment option. However, real-world datasets often violate this assumption, particularly in regions of non-overlap where one treatment group is underrepresented or entirely absent for certain combinations of confounding variables. Traditional approaches, such as trimming and weighting, address these violations but typically modify the target population, potentially introducing bias. The Bayesian Additive Regression Trees with Spline Models (BART+SPL) approach has been proposed as a solution to this issue. BART+SPL combines Bayesian Additive Regression Trees (BART) for imputation in regions of treatment overlap with spline models (SPL) for extrapolation into non-overlap regions, preserving the initial target population. However, BART+SPL's performance is compromised when dealing with high-dimensional covariates. To address this limitation, this paper proposes SBART+SPL, an extension of the BART+SPL framework that integrates SoftBART into the estimation procedure. SoftBART generalizes BART by implementing smooth decision rules and sparsity-inducing splitting probabilities. A simulation study demonstrates that SBART+SPL yields better precision and improved coverage compared to BART+SPL when estimating population average treatment effects (PATE) in the presence of high-dimensional covariates and violations of the positivity assumption. Additionally, the applicability of SBART+SPL is illustrated by re-analyzing an empirical study that evaluates the impact of exposure to natural gas compressor stations on cancer mortality rates across U.S. counties.

Keywords: Bayesian Additive Regression Trees, Causal Inference, Overlap, Extrapolation
JEL-Codes: C01, C11, C31

*The author thanks Christoph Hanck for valuable comments that significantly improved the paper. The paper has been presented at the EUROCIM, 2024, in Copenhagen, and the Workshop on Causal Inference + Machine Learning, 2024, in Groningen. The author is grateful to participants for helpful remarks and interesting discussions

†Chair of Econometrics, Faculty of Business Administration and Economics, University of Duisburg-Essen, Universitätsstraße 12, 45117 Essen, Germany; RGS Econ - Ruhr Graduate School in Economics, Hohenzollernstraße 1-3, 45128 Essen, Germany; Email: lennard.massmann@uni-due.de.

1 Introduction

One of the crucial assumptions to draw causal inference from observational data when using the potential outcome framework is the positivity assumption (Rubin 2005, Hernán & Robins 2020, Li et al. 2022). The assumption postulates that every individual within the considered population possesses a positive probability of receiving either treatment status. Oftentimes, the similar concept of overlap is used to justify the positivity assumption in non-parametric treatment effect estimation by comparing confounder distributions for both treatment groups, the exposure and the control group in the case of a binary treatment. Many real-world datasets suffer from large regions of non-overlap which constitutes a violation of the positivity assumption. Non-overlapping confounder distributions emerge when, at random, no or just a small number of individuals belonging to one treatment group are observed in a specific confounder region (Westreich & Cole 2010). The majority of methods to tackle this issue are based on specific modeling assumptions. Usual approaches like trimming (removing observed data in non-overlap regions) or weighting (reducing the influence of non-overlap observations by decreasing their weights) change the underlying population considered by the initial estimand. For instance, if the initial estimand was the average treatment effect (ATE), those methods are only able to identify the ATE for the trimmed or re-weighted population. Furthermore, approaches like inverse probability weighting (IPW) can lead to weight instability in regions of non-overlap as the estimated probabilities may be close to zero (Crump et al. 2009, Stürmer et al. 2010, Li et al. 2018, Zhu et al. 2021).

Nethery et al. (2019) introduced BART+SPL, a combination of Bayesian Additive Regression Trees (BART) and a spline model (SPL), as an approach to differentiate between observations within overlap and non-overlap regions based on estimated propensity scores. BART+SPL estimates the population average treatment effects (PATE) reliably by tak-

ing into account both of these regions. It is characterized by substantially lower model dependence as well as suitable uncertainty quantification in the non-overlap region via a new Bayesian two-stage procedure. In the imputation phase, BART is used to estimate individual causal effects in the overlap region. In the subsequent smoothing phase, more model dependence is introduced by extrapolating individual causal effects from the overlap region into the non-overlap region relying on a flexible spline model. Simulation studies and an empirical application demonstrate desirable results when faced with lower-dimensional data and different degrees of non-overlap. However, analyzing high-dimensional covariates seems to deteriorate the performance of BART+SPL. If the number of irrelevant covariates that do not influence the determination of potential outcomes rises, bias increases and severe undercoverage issues occur (Nethery et al. 2019).

The contribution of this paper revolves around this issue and is mainly twofold, by proposing a new method that builds upon the framework of BART+SPL. First, instead of using BART for the imputation stage, the SoftBART algorithm introduced by Linero & Yang (2018) is used, as they showed that covariate dimensions can increase nearly exponentially with the sample size for SoftBART. SoftBART is a generalization of BART that uses smooth instead of hard decision rules and sparsity-inducing instead of uniformly distributed splitting probabilities. This enhances the ability to model smoothness in the data generating process, improves uncertainty quantification, and supports selection of relevant covariates by shrinkage of irrelevant covariates. The SBART+SPL approach proposed in this work can be viewed as an extension of BART+SPL to be able to deal with high-dimensional covariates. Second, SBART+SPL's spline model in the smoothing phase does no longer rely on BART+SPL's unidentifiable variance inflation parameter to account properly for higher uncertainty in regions of non-overlap due to SoftBART's ability to estimate this increase in uncertainty when extrapolating from the region of overlap to the region of non-overlap.

The related literature that is concerned with positivity violations when estimating treatment effects in observational studies includes D'Amour et al. (2020) who study different notions of the overlap assumption under high-dimensional covariates in observational studies and derive bounds on imbalances in covariate means when a strict overlap assumption is in place. Moreover, there is recent related work that proposes Bayesian modeling approaches to estimate population average treatment effects while retaining the initial target estimand (Gutman & Rubin 2015, Li et al. 2018, 2019, Nethery et al. 2019, Zhu et al. 2022, Wang et al. 2024).

Gutman & Rubin (2015) propose Multiple Imputation with Two Subclassification Splines (MITSS) to estimate average treatment effects with multiple covariates. They extend the work in Gutman & Rubin (2013) who only consider binary outcomes and one covariate for treatment effect estimation. Gutman & Rubin (2013) and Gutman & Rubin (2015) partition observations into subclasses of units with similar values of X and estimate the PATE by averaging across estimates within these subclasses. Subclasses are related to each other by placing the knots of two regression splines (Wahba 1990) at the boundaries of each subclass where splines are estimated separately for each treatment group considering the distributions of $Y_i(1)$ and $Y_i(0)$ conditional on X_i, respectively. Compared to Gutman & Rubin (2013), MITSS in Gutman & Rubin (2015) allows analyzing continuous outcomes and many covariates. Their simulation study implies that MITSS is generally a valid method and accurate approach, whether X is scalar or multivariate and distributions overlap between treatment and control groups. However, when some units have non-overlapping covariate values across groups, MITSS relies on splines to implicitly extrapolate to regions without observed data for a given treatment which introduces greater bias and improper coverage rates.

By proposing overlap weights, Li et al. (2018, 2019) introduce a novel covariate balancing scheme to estimate average treatment effects reliably. The novel weighting scheme treats

each unit's weight as proportional to the probability of assignment to the opposite treatment group. These overlap weights are bounded and are designed to minimize the asymptotic variance of the weighted average treatment effect within the class of general balancing weights that compensate for differences in the covariate distributions between treatment groups.

Zhu et al. (2022) develop a Bayesian model based on non-parametric Gaussian Process (GP) priors which takes into account the full covariate space and does not need to differentiate between regions of overlap a-priori by incorporating the amount of non-overlap into the GP itself as well as extrapolating with the covariate kernel of the GP.

Wang et al. (2024) apply a local extrapolation method to the Accelerated Bayesian Causal Forest (XBCF) of Krantsevich et al. (2023) by integrating GP into XBCF's leaf nodes, creating a model termed XBCF-GP. This hybrid approach allows for more accurate predictions and better uncertainty quantification for data points outside the training range. Inference on treatment effects is illustrated using simulation data with extreme non-overlap regions wherein either only exposed or unexposed individuals are observed (Wang et al. 2024).

The paper is structured as follows: Section 2 introduces necessary causal inference notation, defines the assumptions of the potential outcome framework and discusses different notions when contrasting between the region of overlap and non-overlap. Section 3 contrasts the conventional BART algorithm of Chipman et al. (2010) with the SoftBART algorithm of Linero & Yang (2018). Section 4 describes the modified two-stage procedure of SBART+SPL based on the BART+SPL algorithm of Nethery et al. (2019). The results of a simulation study focusing on a high-dimensional covariate setup are presented in Section 5. Section 6 compares SBART+SPL to BART+SPL and baseline methods by an empirical analysis of the effect of exposure to natural gas compressor stations on mortality rates in US mid-western counties.

2 Potential outcome framework and region of overlap

Let Y_i^{obs} be the observed continuous[1] outcome of individual i, with individuals $i = 1, ..., n$. The binary treatment status of individual i is defined as $D_i \in \{0, 1\}$ and let $\mathbf{X}_i$ be the p-dimensional vector of confounding values that have been observed for individual i. Consequently, $\mathbf{X}$ is defined as the $n \times p$-dimensional matrix of confounders for all individuals. Using the Stable Unit Treatment Value Assumption (SUTVA), the potential outcome notation is introduced by interpreting the potential outcomes $Y_i(1)$ and $Y_i(0)$ as the values that would have been realized had one observed either treatment status $D_i = 1$ or $D_i = 0$ for individual i. Implicitly, one states with SUTVA that only one potential outcome is realized for individual i such that the non-realized potential outcome is a missing data point. Consequently, one has $Y_i^{obs} = D_i Y_i(1) + (1 - D_i)Y_i(0)$ and $Y_i^{mis} = (1 - D_i)Y_i(1) + D_i Y_i(0)$ with the latter being the missing potential outcome of individual i. That is, one is faced with the fundamental problem of causal inference by having two potential outcomes $Y_i(0)$ and $Y_i(1)$ but only observing one realization Y_i^{obs} of $(Y_i(0), Y_i(1))$. Let us define the following causal effects on different levels of aggregation of an individual (Rubin 2005, Hernán & Robins 2020, Li et al. 2022). Li et al. (2022) define the individual (τ_i), sample average (τ_S), conditional average ($\tau(x)$) and population average (τ_P) treatment effects as follows:

Definition 2.1. Sample and population treatment effects.

$$\tau_i = Y_i(1) - Y_i(0), \tag{2.1}$$

$$\tau_S = \frac{1}{n} \sum_{i=1}^{n} \tau_i, \tag{2.2}$$

$$\tau\left(\mathbf{X}_i\right) = \mathbb{E}[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}], \tag{2.3}$$

$$\tau_P = \mathbb{E}_{\mathbf{X}}[\tau(\mathbf{X}_i)]. \tag{2.4}$$

[1] An extension of the SBART+SPL algorithm in Section 4 that deals with binary outcomes is straightforward. Necessary changes in the setup are described in Appendix D based on Nethery et al. (2019).

To identify τ_P with observational data, usually two additional assumptions next to the above-stated SUTVA are invoked (Li et al. 2022). Unconfoundedness assumes that the treatment assignment is approximately randomized conditional on a sufficiently informative $\mathbf{X}_i$, mimicking a completely randomized controlled trial, such that

$$(Y_i(1), Y_i(0)) \perp D_i | \mathbf{X}_i. \tag{2.5}$$

The positivity assumption postulates that every unit in the considered population possesses a non-zero probability of being assigned to both treatment groups by

$$0 < Pr(D_i = 1 | \mathbf{X}_i) < 1, \tag{2.6}$$

with $p.sc(\mathbf{X}_i) \coloneqq Pr(D_i = 1|\mathbf{X}_i)$ being the propensity score. Unconfoundedness in (2.5) and positivity in (2.6) allow one to view the whole dataset intuitively as a collection of many small $\mathbf{X}_i$-indexed randomized trials. Unconfoundedness ensures conditionally exogenous treatment assignment while positivity ensures that randomization actually occurs in the data. Under (2.5), one receives for the population average treatment effect,

$$\begin{aligned} \tau_P &= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i]] \\ &= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_i(1)|\mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0)|\mathbf{X}_i = \mathbf{x}]] \\ &= \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y^{obs}|\mathbf{X}_i = \mathbf{x}, D_i = 1] - \mathbb{E}[Y^{obs}|\mathbf{X}_i = \mathbf{x}, D_i = 0]]. \end{aligned} \tag{2.7}$$

To identify the conditional expectations in (2.7) non-parametrically, assumption (2.6) is needed. Note that there exists a tension between the assumptions of unconfoundedness and positivity. As the number of covariates increases, the unconfoundedness assumption becomes more plausible but with a rising number of covariates it is also more likely to have sparse data for one or both treatment groups in certain regions of $\mathbf{X}$ (Li et al. 2022).

2.1 Methods to mitigate overlap violation

The positivity assumption in (2.6) can be violated both structurally and randomly. Structural positivity is violated if a unit of interest is not able to obtain treatment such that enlarging the sample size does not alleviate the issue (D'Amour et al. 2020) and will not be the focus of this paper. In contrast, a random positivity violation allows treatment assignment to the unit of interest theoretically, but one cannot (or only rarely) observe this type of treatment empirically within the analyzed data. Increasing the sample size n may mitigate this type of positivity violation. However, increasing the number of covariates p in $\mathbf{X}$ makes the occurrence of a random overlap violation more probable as the likelihood of observing similar units of the exposure and control group for a given covariate combination decreases. This paper deals with the issue of random positivity violations within the potential outcome framework used for causal inference in observational studies for increasing p.

Zhu et al. (2021) point out three main approaches to deal with violations of random positivity in general: trimming, weighting, and extrapolation. The intuition of trimming stems from the idea of identifying and separating out units that seem to violate positivity based on some measure, oftentimes the propensity score, $p.sc(\mathbf{X}_i)$ (Crump et al. 2009, Stürmer et al. 2010). Matching procedures can also be viewed as a form of trimming where treated and controlled units are matched based on their covariate characteristics by finding close distances based on some distance measure between pairs of the exposure and control group (Visconti & Zubizarreta 2018). Random positivity is closely related to the concept of overlap where one defines a region of overlap as a region with covariate values being present in the exposure as well as in the control group. Hence, a region of non-overlap is characterized by covariate values for units that do not exist in a sufficient amount in both groups. Random positivity can be evaluated by assessing overlap through the inspection of

propensity score estimates $\widehat{p.sc}(\mathbf{X}_i)$. Individuals with $\widehat{p.sc}(\mathbf{X}_i) \approx 0.5$ are argued to belong to the overlap region as their occurrence in either treatment group is nearly equally likely. However, with $\widehat{p.sc}(\mathbf{X}_i) \approx 0$ the likelihood of observing a unit with these covariate values in the exposure group is low. Similarly, with $\widehat{p.sc}(\mathbf{X}_i) \approx 1$ the likelihood of observing a unit with these covariate values in the control group is low. For weighting methods like standard inverse propensity weighting (Hernán & Robins 2006, Austin 2011), extreme propensity score estimates might raise the issue of almost infinite weights by dividing by $\widehat{p.sc}(\mathbf{X}_i)$ (Zhu et al. 2021). Furthermore, for either trimming or weighting, the inferential target is relocated from the initial population at hand to the trimmed or matched population to circumvent possible overlap issues.

Using BART+SPL by Nethery et al. (2019) one extrapolates from the regions of overlap to regions of non-overlap. In contrast to trimming or matching, extrapolation based on the distinction between those regions allows statements about treatment effects concerning the original study sample as the initial inferential target population is not relocated.

2.2 Specifying regions of overlap and non-overlap

This paper follows the definition of the region of overlap O and region of non-overlap $O^{\neg}$ of Nethery et al. (2019) by using the estimated propensity scores and assumes that the model for the propensity score is well-specified or the true propensity score is known. Compared to the trimming strategy of Crump et al. (2009), it permits non-overlap regions not only in the tails of the empirical distribution of the estimated propensity scores but also within this distribution. Moreover, the definition in Nethery et al. (2019) may be more tractable when faced with high-dimensional $\mathbf{X}$ in comparison to the direct covariate modeling approach in Hill & Su (2013).

Let o be some point $o \in P.SC = \left[\widehat{p.sc}_{(1)}, \widehat{p.sc}_{(n)}\right] \subset (0,1)$ with $\widehat{p.sc}_{(j)}$ being the j-th order statistic, i.e., $\widehat{p.sc}_{(1)}$ is the minimum of the estimated propensity scores for all

observations $i = 1, ..., n$. Furthermore, let n_d be the number of individuals in each treatment group d and $\widehat{p.sc}_{(i)}$ be the i-th propensity score order statistic in treatment group d.

The point o has a reasonable degree of overlap if one can construct a set,

$$\left\{o, \widehat{p.sc}^d_{(i)}, ..., \widehat{p.sc}^d_{(i+b_O)}\right\}, \tag{2.8}$$

with the following two conditions for the treatment group ($D = 1$) as well as for the control group ($D = 0$), separately:

- The set includes o itself as well as more than b_O estimated propensity scores $\widehat{p.sc}$.
- The set has a range smaller than a_O.

Therefore, one can define the region of overlap O as the set of points o that fulfill those conditions and $O^\neg$ being the complement of this set:

$$O = \left\{ \begin{array}{c} o \in P.SC : \text{range}\left(\left\{o, \widehat{p.sc}^d_{(i)}, ..., \widehat{p.sc}^d_{(i+b_O)}\right\}\right) < a_O, \\ \text{for some } i = 1, ..., n_d - b_O, \\ \text{for } d = \{0, 1\}. \end{array} \right\}. \tag{2.9}$$

Any user of this overlap definition is left with choosing appropriate values for a_O and b_O. As outlined in Nethery et al. (2019), the tuning parameters have recommended default choices with $a_O = 0.1$ and $b_O = 7$, respectively. This implies that a point o belongs to the region of overlap if, for each treatment group, one can construct a set with 7 estimated propensity scores around this point such that this set has a range that is smaller than 0.1. In general, lower values of a_O together with higher values of b_O define the region of non-overlap more conservatively. Estimated propensity scores belonging to the set are allowed to only vary mildly by opting for a low a_O and one needs relatively many of these estimated propensity scores that are close to o due to the large b_O. Likewise, high values for a_O and low values for b_O define the region of non-overlap less conservatively.

Given these region-of-overlap and region-of-non-overlap definitions, BART+SPL in Nethery et al. (2019) and SBART+SPL as its proposed extension in Section 4 use observations in O to extrapolate treatment effect patterns into $O^{\neg}$ using BART or SoftBART, respectively. The following Section 3 explains BART and motivates the usage of SoftBART as a suitable extension of BART especially in sets with high-dimensional covariates.

3 From BART to SoftBART

This section mainly follows the notation of Linero & Yang (2018) to describe the BART and SoftBART algorithms. Let us consider the general semiparametric Gaussian regression problem with

$$Y_i = f_0(\mathbf{X}_i) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{3.1}$$

to exemplify how BART is used for predicting outcome Y_i given $\mathbf{X}_i$. In Section 4, the missing potential outcome in the SBART+SPL algorithm is imputed by (4.3) which targets a regression problem similar to (3.1). Let us define the non-parametric regression function with $f_0(\mathbf{X}_i = x)$ being some realization of the function

$$\begin{aligned} f(\mathbf{X}_i = x) &= \sum_{j=1}^{J} g\left(x; \mathcal{T}_j, \mathcal{M}_j\right) \\ &= \sum_{j=1}^{J} \sum_{l=1}^{L_j} \mu_{jl} \phi\left(x; \mathcal{T}_j, l\right), x \in \mathbb{R}^p, \end{aligned} \tag{3.2}$$

with $\mathcal{T}_j$ the branching process of tree $j \in \{1, ..., J\}$ while $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$ represents the leaf node parameters of tree j with number of leaf $l \in \{1, ..., L_j\}$. For each tree j and given data x, the function $g\left(x; \mathcal{T}_j, \mathcal{M}_j\right)$ relates a branching process to the leaf node parameters and can be reformulated by multiplying each leaf node parameter μ_{jl}

with the respective splitting rule $\phi(x; \mathcal{T}_j, l)$. The hard splitting rule of BART is further described in (3.4). BART uses the sum-of-trees specification in (3.2) to learn flexible functions $g(x; \mathcal{T}_j, \mathcal{M}_j)$ that are able to predict Y_i based on $\mathbf{X_i}$ and prior distributions on $\mathcal{T}_j, \mathcal{M}_j$ and error varianceσ^2 from (3.1). These prior distributions are discussed in more detail in Subsections 3.1.1 for BART, 3.2 for Dirichlet Additive Regression Tress (DART), and 3.3 for SoftBART.

One can view the BART algorithm of Chipman et al. (1998, 2006, 2010) as an aggregation of weak-learning decision trees as described in the decision tree boosting setup, but in a Bayesian fashion. The method of gradient boosting machines (GBM) is a statistical framework that can be used to aggregate many decision trees for a robust predictive performance. The GBM algorithm searches for an additive model with minimal residual loss. Based on gradient calculations and residual model fitting for loss minimization, the current model is added to the former model repeatedly (Friedman 2001). GBM can be compared to other tree aggregation approaches like bootstrap aggregation, also known as bagging, or random forests (James et al. 2021). Bagging produces decision tree ensembles by constructing trees independently on bootstrap samples (Breiman 1996). Random forests improves on possible shortcomings of bagging (overly similar decision trees leading to insufficient model space exploration) by decorrelating the decision trees (Breiman 2001). That is, trees are still constructed on independent bootstrap samples but only using a random covariate subset for each tree such that bagging can be expressed as a special case of random forest. In contrast to bagging and random forests, GBM does not use bootstrap samples but just the actual observed data. Moreover, trees are constructed dependently, based on the residual of the former fitted tree.

The BART algorithm proceeds similar to GBM in that it also only uses the observed data and tree construction is performed dependently. Different from GBM, BART does not fit an entirely new tree to the residuals of the former tree. Instead, BART has a built-in

perturbation process that changes the tree structure of the previous tree by either adding or pruning branches or modifying each of the leaf node predictions. This prevents getting stuck in local minima via an increased exploration of the model space. To avoid overfitting, GBM uses tuning parameters by limiting the maximum depth of each tree resulting in weak learning trees and scaling down the individual contribution of each new tree. In contrast, BART reduces overfitting in a rather data-driven way by using a prior distribution for each tree to regularize its size and fit. The independent tree priors prefer the construction of rather small trees and lead to leaf parameter values approaching zero (Hill et al. 2020).

Besides its flexible nonparametric function estimation approach, BART possesses further advantages that are useful for drawing inference in empirical studies with observational data. First, the BART algorithm shows good performance results in setups with high noise that are prevalent in the causal inference literature of economics or medicine. Competitive and superior performance compared to other methods has been demonstrated in data challenges and empirical applications (Dorie et al. 2019, Hu et al. 2020, Wendling et al. 2018). Second, the prior distributions used in the BART algorithm favor data patterns with low-order interactions over high-order interactions. Although high-order interactions play a central role in pattern recognition for, i.e., languages or images, they are argued to be of limited relevance in the traditional statistical analysis of the social sciences. In these circumstances of low-order interactions, BART approaches optimal posterior concentration rates (Rockova & van der Pas 2019, Saha 2023).

A drawback of the BART algorithm is its reliance on step-wise continuous functions that lead to non-smooth predictions. The SoftBART algorithm of Linero & Yang (2018) presented in Section 3.3 has been proposed to resolve this shortcoming by introducing smoothness into the conventional BART algorithm of Chipman et al. (1998, 2010). Linero & Yang (2018) and Rockova & van der Pas (2019) show that the convergence rate of BART can be improved via smoothness. Moreover, they derive suitable posterior concentration

rates for SoftBART when the number of relevant predictors p_{true} is much smaller than the number of variables p and the true data-generating function is driven by low-order interactions. Recently, the SoftBART package (Linero 2022) has been introduced to flexibly implement BART models that can adapt to sparsity and smoothness. This package is used to program the SBART+SPL algorithm in Section 4 and apply it in the simulation study in Section 5 and the empirical application in Section 6. The following Subsection 3.1 further presents the standard notation of BART before Subsection 3.1.1 describes the main characteristics of the corresponding prior distributions. Subsection 3.2 generalizes BART to DART by allowing for sparsity-inducing splitting probabilities while Subsection 3.3 generalizes DART to SoftBART by allowing for smoothness-inducing decision rules.

3.1 Bayesian Additive Regression Trees

The standard BART algorithm by Chipman et al. (2010) describes an ensemble of decision trees with $(\mathcal{T}_j, \mathcal{M}_j) \overset{i.i.d.}{\sim} \pi_{\mathcal{T}}(\mathcal{T}_j)\,\pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j)$ with $(\pi_{\mathcal{T}}, \pi_{\mathcal{M}})$ being parameter priors of the decision trees. Referring back to the semiparametric Gaussian problem in (3.1), BART takes the decision trees j to be independent such that one can write

$$\pi\left((\mathcal{T}_1, \mathcal{M}_1), ..., (\mathcal{T}_J, \mathcal{M}_J), \sigma|\theta\right) = \left[\prod_{j=1}^{J} \pi_{\mathcal{T}}(\mathcal{T}_j|\theta)\,\pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j)\right]\pi_{\sigma}(\sigma) \tag{3.3}$$

with $\pi_{\mathcal{M}}(\mathcal{M}_j|\mathcal{T}_j) = \prod_{l=1}^{L_j} \pi_{\mathcal{M}}(\mu_{jl}|\mathcal{T}_j)$ and θ being the vector of hyperparameters specified below. Consequently, one needs to define prior distributions for $(\pi_{\mathcal{T}}, \pi_{\mathcal{M}}, \pi_{\sigma})$. The prior on $\pi_{\mathcal{T}}$ is twofold, focusing on (1) the tree shape and (2) the splitting rules connected to each branch node. The prior on $\pi_{\mathcal{M}}$ describes the distribution of leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$. The prior on π_{σ} is a distribution for the noise variance.

3.1.1 Prior specification

Following Chipman et al. (2010), the prior on the tree shape of tree $\mathcal{T}_j$ can be outlined as a branching process. Each tree begins with a root node of depth $d = 0$. Then, it is iteratively decided how deep one is growing the tree. The root node is converted into a branch node b with two child nodes with probability $Pr\left(b \text{ is branch node}\right) = \frac{\gamma}{(1+d)^{\beta}}$ and is converted into a leaf node with its complementary probability. This procedure repeats until all nodes at a given level of depth level are leaf nodes. Consequently, one has to specify the hyperparameters (γ, β). This paper uses the default values of Chipman et al. (2010) by $\gamma = 0.95$ and $\beta = 2$.

The prior on the splitting rules connected to each branch node b is defined by sampling a predictor j_b and a cutpoint C_b. Based on those sampled values, one can compare x_{j_b} and C_b to decide how to channel x down the tree. The standard BART algorithm samples j_b and C_b from separate uniform distributions. The hard branch splitting rules of the standard BART algorithm can be formalized by specifying $\phi\left(x; \mathcal{T}_j, l\right)$ used in (3.2) by

$$\phi\left(x; \mathcal{T}_j, l\right) = \prod_{b \in \mathcal{A}(l)} \mathbb{1}\left(x_{j_b} \leq C_b\right)^{R_b} \mathbb{1}\left(x_{j_b} > C_b\right)^{1-R_b}. \tag{3.4}$$

Let us denote $\mathcal{A}(l)$ as nodes that are ancestral to leaf node l. Furthermore, $R_b \in \{0, 1\}$, such that $R_b = 0$ if the path from the root to l goes right at b and $R_b = 1$ if left.

For the prior on the leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$, one can exploit conjugacy of the normal distribution $\mathcal{N}(\mu_\mu, \sigma_\mu^2)$ to marginalize out μ_{jl} as in Chipman et al. (2010). For the standard BART algorithm, the outcome variable Y is transformed to be placed into the interval $[-0.5, 0.5]$ and one samples $\mu_{jl} \sim \mathcal{N}(\mu_\mu = 0, \sigma_\mu^2 = 0.5/k\sqrt{J})$. Consequently, one has to specify two further hyperparameters (k, J) with $k = 2$ and $J = 200$ being default values for the standard BART algorithm.

Again, one can exploit conjugacy for the prior on the noise variance σ^2. For $\pi_\sigma(\sigma)$ one

can use the inverse chi-square distribution such that $\sigma^2 \sim \upsilon\lambda/\chi^2_\upsilon$ with default choices of $(\upsilon = 3, q_\lambda = 0.9)$ to avoid neither an overly conservative nor an excessively aggressive prior. Here, λ is chosen such that the the probability of σ being lower than the initial $\hat{\sigma}$ is q_λ by determining $Pr\left(\sigma < \hat{\sigma}\right) = q_\lambda$. The naive $\hat{\sigma}$ is either estimated by being the sample standard deviation of Y or the residual standard deviation from a simple linear regression of Y on $\mathbf{X}$. This ensures that the prior is weakly informative and favors values of σ that are similar to its naive estimate, but still allows for variation. Oftentimes, the random draw is taken from the related inverse-gamma distribution with $\sigma^2 \sim \text{IG}\left(\frac{\upsilon}{2}, \frac{\upsilon\lambda}{2}\right)$.

3.1.2 Posterior inference

Algorithm 1 describes the pseudo-code of one iteration of the general Bayesian backfitting procedure to update $(\mathcal{T}_j, \mathcal{M}_j)$ for BART (Chipman et al. 2010, Hill et al. 2020). Bayesian backfitting adopts a blocked Metropolis-Hastings strategy to explore the posterior distribution (Chipman et al. 1998, Wu et al. 2007). In essence, the sampling proceeds in a structured manner: the tree $\mathcal{T}_j$ is first drawn from its marginal posterior distribution, followed by drawing the leafs $\mathcal{M}_j$ from their full conditional distribution. More complicated models use a Metropolis-within-Gibbs Markov Chain Monte Carlo model to update other parameters by exploiting further Gibbs or Metropolis-Hastings steps as presented in Algorithm 2 later on for SBART+SPL.

In step 1, Algorithm 1 proposes a new tree structure $\mathcal{T}_j^\star$ based on some proposal distribution based on the current tree structure $\mathcal{T}_j$ by $q(\mathcal{T}_j^\star; \mathcal{T}_j)$. Mostly, BART implementations rely on random perturbations of the current tree structure by one of the following moves: *Grow/Birth* turns a considered leaf node into a branching node and splits into two new leaf nodes, *Prune/Death* collapses two neighboring leaf nodes back to one leaf node, *Change* modifies the decision rule associated with a non-terminal node, and, *Swap* exchanges the decision rules of two non-terminal nodes (Hill et al. 2020, Linero & Yang 2018). Wu et al.

(2007) and Pratola (2016) provide an extended discussion of alternative proposal distributions within BART.

The second step sets the Metropolis ratio, a_{MR} in Equation (3.6), by using the proposal distribution from step 1 and computing the integrated likelihood function,

$$\mathcal{L}(\mathcal{T}_j; T_{(j)}, M_{(j)}, \theta) = \int \left(\prod_{i=1}^{n} p(Y_i \mid \mathcal{T}_h, M_j, T_{(j)}, M_{(j)}, \theta) \right) p(M_j \mid \mathcal{T}_j, \theta)\, dM_j, \tag{3.5}$$

where $T_{(j)} \equiv \{T_{\mathcal{J}} : 1 \leq \mathcal{J} \leq J,\ \mathcal{J} \neq j\}$ is the set of all tree structures except T_j while the set $M_{(j)}$ is defined similarly for leaf parameters (Hill et al. 2020). Moreover, the vector θ collects all necessary parameters considered in the branching process in Section 3.1.1 and might be extended to include further parameters in more involved models (i.e. bandwidth parameter τ^{bw} in Section 3.3 for SoftBART).

Step 3 accepts the tree proposal $\mathcal{T}_j^{\star}$ with probability $\min(1, a_{MR})$. If the tree proposal is rejected, T_j stays at its current status. Finally, leaf parameters $\mathcal{M}_j$ are drawn from their full conditional distribution in step 4.

Algorithm 1: One iteration of Bayesian backfitting for BART to update $(\mathcal{T}_j, \mathcal{M}_j)$

Input :

- Initial parameters $\left\{\mathcal{T}_j^{(0)}, \mathcal{M}_j^{(0)}\right\}$
- Hyperparameters $\theta = (\gamma, \beta, k, J)$ as specified in Sections 3.1.1
- Observed data $\mathbf{Y}^{obs}, \mathbf{X}$

Output: Updated values for $\left(\mathcal{T}_{|}, \mathcal{M}_{|}\right)$

for $j \leftarrow 1$ **to** J **do**

1. Propose new tree structure $\mathcal{T}_j^{\star}$ from proposal distribution $q(\mathcal{T}_j^{\star}; \mathcal{T}_j)$
2. Set Metropolis ratio, a_{MR}, to

$$a_{MR} \leftarrow \frac{\mathcal{L}(\mathcal{T}_j^{\star}; \mathcal{M}_j, \theta) p(\mathcal{T}_j^{\star})}{\mathcal{L}(\mathcal{T}_j; \mathcal{M}_j, \theta) p(\mathcal{T}_j)} \frac{q(\mathcal{T}_j; \mathcal{T}_j^{\star})}{q(\mathcal{T}_j^{\star}; \mathcal{T}_j)} \tag{3.6}$$

3. Set $\mathcal{T}_j \leftarrow \mathcal{T}_j^{\star}$ with probability $\min(1, a_{MR})$
4. Sample $\mathcal{M}_j \sim p(\mathcal{M}_j | \mathcal{T}_j, \mathcal{T}_{(j)}, \mathcal{M}_{(j)}, \theta, \mathbf{Y}^{obs}, \mathbf{X})$

end

3.2 Dirichlet Additive Regression Trees (DART)

The Dirichlet Additive Regression Trees (DART) algorithm extends the BART algorithm with regard to the prior on the splitting rules of branch node b by adding a sparsity-inducing prior (Linero 2018). It samples $j_b \sim Categorical(s)$ with probability vector $s = (s_1, ..., s_p)^\top$. Therefore, one has to specify an additional prior on s which induces sparsity into the splitting probabilities. That is, we sample

$$s \sim \text{Dirichlet}\left(\alpha/p^\xi, ..., \alpha/p^\xi\right) \tag{3.7}$$

with $\frac{\alpha}{\alpha+p} \sim \text{Beta}\left(\alpha_s = 0.5, \beta_s = 1\right)$ and $\xi = 1$ such that α governs the assumed sparsity in f with (α_s, β_s) balancing between sparse and non-sparse setups. Second, the cutpoints are sampled by $C_b \sim \text{Uniform}(a_u, b_u)$ with (a_u, b_u) chosen such that the cell $\mathbb{R}^p$ is split along the j_b-th coordinate.

3.3 Soft Bayesian Additive Regression Trees (SoftBART)

The SoftBART algorithm builds upon the DART algorithm. It extends the procedure by implementing smoothness-inducing decision rules (Linero & Yang 2018, Linero 2022). The algorithm replaces the hard decision rules $\mathbb{1}\left(x_{j_b} \leq C_b\right)$ and $\mathbb{1}\left(x_{j_b} > C_b\right)$ as in 3.4 with the smooth decision rule $\psi\left(\frac{x_{j_b} - C_b}{\tau_j^{bw}}\right)$ leading to

$$\phi\left(x; \mathcal{T}_j, l\right) = \prod_{b \in \mathcal{A}(l)} \psi\left(\frac{x_{j_b} - C_b}{\tau_j^{bw}}\right)^{R_b} \left[1 - \psi\left(\frac{x_{j_b} - C_b}{\tau_j^{bw}}\right)\right]^{1-R_b}. \tag{3.8}$$

The SoftBART algorithm assigns each tree $\mathcal{T}_j$ a separate bandwidth parameter $\tau_j^{bw} \sim \text{Exp}\left(\text{scale} = 0.1\right)$.[2] Moreover, $\psi\left(\frac{x_{j_b} - C_b}{\tau_j^{bw}}\right)$ is computed using a logistic function such that $\psi(\frac{x_{j_b} - C_b}{\tau_j^{bw}}) = \left(1 + \exp\left(-\frac{x_{j_b} - C_b}{\tau_j^{bw}}\right)\right)^{-1}$. Using τ_j^{bw} in that way provides many possible shapes of the gating functions ψ allowing for approximates of step functions besides smooth

[2] Note that if $\tau_j^{bw} \to 0$ yields a standard decision tree.

functions. Added to that and with regard to the prior on the leaf node parameters $\mathcal{M}_j = \left(\mu_{j1}, ..., \mu_{jL_j}\right)$, the SoftBART algorithm places an additional hyperprior on σ_μ^2 by sampling $\sigma_\mu \sim \text{Cauchy}^+(\hat{\sigma}_\mu)$ with $\hat{\sigma}_\mu = 0.5/k\sqrt{J}$ and choosing by default fewer trees with $J = 20$.

The differences between SoftBART and BART are illustrated in two toy examples that can be found in Appendix A and B and which are closely related to illustrations presented in Hahn et al. (2020) and Linero & Yang (2018). The first example points out the difference between hard and soft decision rules in a simple prediction task of a smooth sine curve as well as a step function and indicate lower RMSE for SoftBART compared to BART as BART leads to non-smooth jumps around the true data-generating function. The second example shows an improvement in uncertainty quantification in the region of non-overlap in a simple conditional average treatment effect estimation setup when estimating the response surfaces of the exposure and the control group with SoftBART instead of BART. BART does not account for an increase in uncertainty and predicts a constant treatment effect in the region of non-overlap. Instead, SoftBART inflates the variance around treatment effects more properly and allows for a more flexible treatment effect estimates when faced with weak overlap. The SoftBART properties of improved predictive power and enhanced uncertainty quantification are leveraged within the SBART+SPL algorithm described in Section 4 to alleviate the shortcomings of BART+SPL concerning precision and coverage for population average treatment effect estimation in the presence of sparse data.

4 SBART+SPL

This section describes the procedure to generate draws from the posterior density of the PATE using SBART+SPL. A simplified pseudo-code for the Markov Chain Monte Carlo (MCMC) algorithm in the style of Nethery et al. (2019) can be found in Algorithm 2. The

main differences compared to the original BART+SPL algorithm are mainly two-fold. In the imputation phase within the region of overlap (steps 1 to 5 in Algorithm 2), SoftBART (Linero & Yang 2018, Linero 2022) replaces BART (Chipman et al. 2010) for Bayesian backfitting such that potential sparsity in the data is handled adequately by using sparsity-inducing splitting probabilities. In the extrapolation phase concerning the region of non-overlap (steps 6 to 9 in Algorithm 2), SBART+SPL does not rely on the introduced variance inflation parameter in BART+SPL for the spline extrapolation. As indicated in Section 3.3, SoftBART improves uncertainty quantification reasonably compared to BART such that one can avoid the use of this unidentifiable variance inflation parameter. Computationally, this paper implements the SBART+SPL algorithm by embedding the SoftBART algorithm into the Bayesian backfitting algorithm of BART+SPL using the `MakeForest()` function and its corresponding `Rcpp_Forest` data structure as outlined in Linero (2022) for the `R` programming language (R Core Team 2021).

4.1 Imputation using SoftBART

Steps 1 to 3 of the algorithm perform the Bayesian backfitting for the SoftBART algorithm as described in Linero & Yang (2018) using observed data from the region of overlap, $\mathbf{Y}_O^{obs}$, where the region of overlap is specified using the approach of Nethery et al. (2019) described in Section 2.2. Note that the original posterior distribution for a given tree of the SoftBART algorithm,

$$\pi\left(\mathcal{T}_j, \mathcal{M}_j | \mathbf{Y}_O^{obs}, \sigma^2_{Imp}, \tau_j^{bw}, \mathcal{T}_j, ..., \mathcal{T}_J, \mathcal{M}_j, ..., \mathcal{M}_J\right), \tag{4.1}$$

can be written more compactly using partial residuals. Let us define $\mathbf{Y}_O^{obs} = [Y_1^{obs}, ..., Y_{n_O}^{obs}]'$ as the vector of observed outcomes of individuals o in the region of overlap as defined in Section 2.2 and n_O the number of observations in the region of overlap. Let the partial

residuals be

$$\mathbf{V}_{Oj} = \mathbf{Y}_O^{obs} - \sum_{v \neq j} g(\mathbf{X}_O; \mathcal{T}_v, \mathcal{M}_v). \tag{4.2}$$

Then, $\mathbf{V}_{Oj}$ allows for Bayesian backfitting where all remaining tree parameters are held constant while drawing from one particular tree. Step 1 mainly updates the tree structures $\mathcal{T}_j$ and the leaf parameters $\mathcal{M}_j$ for each tree j, similarly as discussed in Subsection 3.1.1 for the BART algorithm. Updating $\mathcal{T}_j$ and τ_j^{bw} is solved via Metropolis-Hastings. In contrast to BART+SPL, the update for the bandwidth parameters τ_j^{bw} is necessary in SBART+SPL to allow for flexible gating function shapes as outlined in Subsection 3.3. The necessary marginalization over $\mathcal{M}_j$ is performed in closed form due to the Gaussian error assumption. Choosing the number of trees $J = 20$ to be much smaller than in the standard BART algorithm (there we have $J = 200$) balances the increase in computational time due to this marginalization step. $\mathcal{M}_j$ can be drawn from a normal distribution. After iterating over all J trees, the splitting probabilities in parameter s, defined in 3.7, are updated via Gibbs sampling using a Dirichlet distribution as indicated in step 2. In step 3, one draws the noise variance σ_{Imp}^2 and leaf variance parameters σ_μ^2 from Inverse-Gamma distributions, as already discussed in Subsection 3.1.1 and Subsection 3.3. Moreover, the sparsity-control parameter α is retrieved using slice sampling by Neal (2003). Therefore, steps 1 to 3 are just a replacement of the BART algorithm with the SoftBart algorithm in the original BART+SPL algorithm of Nethery et al. (2019) to perform a smoothed imputation in step 4.

Similarly to the notation from above for $\mathbf{Y}_O^{obs}$, the corresponding vector of missing outcomes in the region of overlap is $\mathbf{Y}_O^{mis} = [Y_1^{mis}, ..., Y_{n_O}^{mis}]'$. As discussed in Section 2 related to SUTVA, only one potential outcome is realized for each individual i. For individual o in the overlap region, this non-realized potential outcome is captured with $\mathbf{Y}_O^{mis}$ as a missing data point and needs to be imputed. The vector $\tilde{\mathbf{Y}}_O^{mis}$ collects the draws

of the imputed values of $\mathbf{Y}_O^{mis}$ from its posterior predictive distribution. More precisely, a specific Y_o^{mis} is imputed in step 4 by applying SoftBART to

$$Y_o^{obs} = \sum_j^J g\left(D_o, \widehat{p.sc}_o, \mathbf{X}_o; \mathcal{T}_j, \mathcal{M}_j\right) + \epsilon_o, \epsilon_o \sim \mathcal{N}(0, \sigma_{Imp}^2), \tag{4.3}$$

similar to the usage in Equations (3.2) and (3.1). Here, the estimated propensity score $\widehat{p.sc}_o, \mathbf{X}_o$ serves as another predictor while D_o lets us differentiate between treatment and control group. Let us collect parameters for the imputation phase in $\theta^{Imp} = \{\sigma_{Imp}^2, \mathcal{T}_j, ..., \mathcal{T}_J, \mathcal{M}_j, ..., \mathcal{M}_J\}$. This gives us a posterior distribution of $p\left(\theta^{Imp}|\mathbf{Y}_O^{obs}\right)$ which can be used to obtain draws from the posterior predictive distribution

$$p\left(\mathbf{Y}_O^{mis}|\mathbf{Y}_O^{obs}\right) = \int p\left(\mathbf{Y}_O^{mis}|\mathbf{Y}_O^{obs}, \theta^{Imp}\right) p\left(\theta^{Imp}|\mathbf{Y}_O^{obs}\right) \mathrm{d}\theta^{Imp}. \tag{4.4}$$

These draws are defined as imputed values of missing potential outcomes in the region of overlap, $\tilde{\mathbf{Y}}_O^{mis}$. Subsequently, the vector of estimated individual treatment effects for the overlap-region O, $\tilde{\boldsymbol{\Delta}}_O = [\tilde{\Delta}_1, ..., \tilde{\Delta}_{n_O}]'$, is computed in step 5 of Algorithm 2 such that the individual treatment effect of a specific $o \in \{1, ..., n_O\}$ are obtained as

$$\tilde{\Delta}_o = \begin{cases} Y_o^{obs} - Y_o^{mis} \text{ , if } D_o = 1 \\ Y_o^{mis} - Y_o^{obs} \text{ , if } D_o = 0 \end{cases}. \tag{4.5}$$

Algorithm 2: SBART+SPL

Input :

- Initial parameters $\left\{\mathcal{T}_j^{(0)}, \mathcal{M}_j^{(0)}, {\sigma^2_{Imp}}^{(0)}, {\sigma^2_{\mu}}^{(0)}, s^{(0)}, \alpha^{(0)}\boldsymbol{\beta}_{Smo}^{(0)}, {\sigma^2_{Smo}}^{(0)}\right\}$
- Hyperparameters $\theta = (\gamma, \beta, \alpha_s, \beta_s, k, J, \upsilon, q_\lambda,)$ as specified in Sections 3.1.1, 3.2, 3.3
- Observed data $\mathbf{Y}^{obs}, \mathbf{X}$ and region of overlap/non-overlap $O, O^{\neg}$

Output: $\tilde{\Delta}_P$ as m draws of the population average treatment effect from the posterior density

for $m \leftarrow 1$ **to** M **do**

1. **for** $j \leftarrow 1$ **to** J **do**
   - 1.1 Draw $\mathcal{T}_j^{(m)}$ from $p\left(\mathcal{T}_j|(\tau_j^{bw})^{(m)}, \mathbf{V}_{Oj}^{(m-1)}, {\sigma^2_{Imp}}^{(m-1)}\right)$ using the Metropolis Hastings algorithm described by Chipman et al. (1998)
   - 1.2 Draw ${\tau_j^{bw}}^{(m)}$ from $p\left(\tau_j^{bw}|\mathcal{T}_j^{(m)}, \mathbf{V}_{Oj}^{(m-1)}, {\sigma^2_{Imp}}^{(m-1)}\right)$ using the Metropolis Hastings algorithm described by Chipman et al. (1998)
   - 1.3 Draw $\mathcal{M}_j^{(m)}$ from $p\left(\mathcal{M}_j|\mathbf{V}_{Oj}^{(m-1)}, {\sigma^2_{Imp}}^{(m-1)}, \mathcal{T}_j^{(m)}\right)$ through a random sample from the Normal distribution

   end
2. Draw $s^{(m)}$ through a random draw from a Dirichlet distribution
3. Draw noise variance ${\sigma^2_{Imp}}^{(m)}$ and leaf variance parameters ${\sigma^2_{\mu}}^{(m)}$ from Inverse-Gamma distributions and the sparsity-control parameter $a^{(m)}$ using slice sampling
4. Draw ${\mathbf{Y}_O^{mis}}^{(m)}$ from $p\left(\mathbf{Y}_O^{mis}|\mathbf{Y}_O^{obs}, \mathcal{T}_j^{(1)}, ..., \mathcal{T}_j^{(J)}, \mathcal{M}_j^{(1)}, ..., \mathcal{M}_j^{(J)}, {\sigma^2_{Imp}}^{(m)}\right)$ through a random sample from a Normal distribution
5. Form $\tilde{\Delta}_O^{(m)}$ as a linear combination of ${\mathbf{Y}_O^{mis}}^{(m)}, \mathbf{Y}_O^{obs}$
6. Draw $\boldsymbol{\beta}_{Smo}^{(m)}$ from $p\left(\boldsymbol{\beta}_{Smo}|\tilde{\Delta}_O^{(m)}, {\sigma^2_{Smo}}^{(m-1)}\right)$
7. Draw ${\sigma^2_{Smo}}^{(m)}$ from $p\left(\sigma^2_{Smo}|\tilde{\Delta}_O^{(m)}, \boldsymbol{\beta}_{Smo}^{(m)}\right)$ through a random sample from an Inverse-Gamma distribution
8. $\tilde{\Delta}_{O^{\neg}}^{(m)}$ from $p\left(\tilde{\Delta}_{O^{\neg}}|\tilde{\Delta}_O^{(m)}, \boldsymbol{\beta}_{Smo}^{(m)}, {\sigma^2_{Smo}}^{(m)}\right)$ through a random sample from a Normal distribution
9. Draw $\hat{\Delta}_P^{(m)}$ by executing B iterations of the Bayesian bootstrap on $\left\{\tilde{\Delta}_O^{(m)}, \tilde{\Delta}_{O^{\neg}}^{(m)}\right\}$ and randomly selecting one of the B bootstrap sample averages

end

4.2 Extrapolation without variance inflation parameter

Steps 6 to 9 are then similar steps as in BART+SPL except that the tuning parameter for variance inflation in the non-overlap region drops out. The proposed smoothing procedure in Nethery et al. (2019) estimates individual treatment effects based on the estimates retrieved from the imputation phase and the identified patterns in the region of overlap.

Let $\mathbf{Y}_{O}^{mis}$ and $\mathbf{Y}_{O^{\neg}}^{mis}$ be the vectors of missing potential outcomes of individuals in the region of overlap and non-overlap, respectively. The two smoothing models use restricted cubic splines[3] $rcs(z)$ with basis z to extrapolate individual treatment effects into the region of non-overlap and read as follows

$$\tilde{\Delta}_o = \mathbf{W}_o' \boldsymbol{\beta}_{Smo} + \epsilon_o \ , \ \epsilon_o \sim \mathcal{N}\left(0, \sigma^2_{Smo}\right). \tag{4.6}$$

Note that σ^2_{Smo} is the noise variance for the smoothing model which can be different from the noise variance in Equation (4.3) for the imputation model. Moreover, $\tilde{\Delta}_o$ relates to an individual treatment effect for the overlap region already computed during the imputation phase as in Equation (4.5). Moreover, $\mathbf{W}_o$ is defined by

$$\mathbf{W}_o = \begin{cases} \left[rcs\left(\widehat{p.sc}_o\right) \ , \ rcs\left(Y_o^{\star}(1)\right) \ , \ \mathbf{X}_o\right]', \ \text{if } D_o = 1 \\ \left[rcs\left(\widehat{p.sc}_o\right) \ , \ rcs\left(Y_o^{\star}(0)\right) \ , \ \mathbf{X}_o\right]', \ \text{if } D_o = 0 \end{cases}. \tag{4.7}$$

With Equation (4.7), two smoothing models are defined: One for treated individuals in the non-overlap region ($D_{o^{\neg}} = 1$) and one for non-treated individuals in the non-overlap

[3]Knots are chosen by the $(.1, .25, .5, .75, .9)$-quantiles of the estimated propensity score values and $(.2, .4, .6, .8)$-quantiles of the $Y_o^{\star}$ values.

region ($D_{o^\neg} = 0$). For the former case, one uses

$$Y_o^\star(1) = \begin{cases} Y_o^{obs} \text{ , if } D_o = 1, \\ \tilde{Y}_o^{mis} \text{ , if } D_o = 0 \end{cases}, \tag{4.8}$$

with $\tilde{Y}_o^{mis}$ as in Equation (4.3). For the latter case, one similarly uses

$$Y_o^\star(0) = \begin{cases} Y_o^{obs} \text{ , if } D_o = 0, \\ \tilde{Y}_o^{mis} \text{ , if } D_o = 1 \end{cases}. \tag{4.9}$$

Equation (4.6) constitutes a linear regression model such that one can retrieve updates of $(\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo})$ by drawing from Normal and conjugate Inverse-Gamma distributions for each of the two smoothing models, separately. With the updates of these spline parameters $\theta^{Smo} = \{\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}\}$, one can draw from the posterior predictive distribution for the non-overlap region:

$$p\left(\tilde{\Delta}_{o^\neg} | \tilde{\Delta}_o\right) = \int p\left(\tilde{\Delta}_{o^\neg} | \mathbf{W}^\star_{o^\neg}, \theta^{Smo}\right) p\left(\theta^{Smo} | \tilde{\Delta}_o\right) \mathrm{d}\theta^{Smo}, \tag{4.10}$$

where the likelihood model reads

$$p\left(\tilde{\Delta}_{o^\neg} | \mathbf{W}^\star_{o^\neg}, \theta^{Smo}\right) \sim \mathcal{N}(\mathbf{W}^{\star\prime}_{o^\neg} \boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}) \tag{4.11}$$

This draw is a sample from the normal distribution with mean of $\mathbf{W}^{\star\prime}_{O^\neg} \boldsymbol{\beta}_{Smo}$ and variance of σ^2_{Smo}, where $\mathbf{W}^\star_{o^\neg}$ consists of

$$\mathbf{W}^\star_{o^\neg} = \begin{cases} [rcs\left(\widehat{p.sc}_{o^\neg}\right) \text{ , } rcs\left(Y^\star_{o^\neg}(1)\right) \text{ , } \mathbf{X}_{o^\neg}]', \text{ if } D_{o^\neg} = 1 \\ [rcs\left(\widehat{p.sc}_{o^\neg}\right) \text{ , } rcs\left(Y^\star_{o^\neg}(0)\right) \text{ , } \mathbf{X}_{o^\neg}]', \text{ if } D_{o^\neg} = 0 \end{cases}, \tag{4.12}$$

and $(\boldsymbol{\beta}_{Smo}, \sigma^2_{Smo})$ generated from either of the two smoothing models described above. Hence, an extrapolated individual treatment effect for the non-overlap region, $\tilde{\Delta}_{o^\neg}$, is sampled from:

$$p\left(\tilde{\Delta}_{O^\neg}|\mathbf{W}^\star, \theta^{Smo}\right) \sim \mathcal{N}(\mathbf{W}^\star \boldsymbol{\beta}_{Smo}, \sigma^2_{Smo}), \quad \text{where } (\boldsymbol{\beta}, \sigma^2_S) \sim p(\theta^{Smo} \mid \tilde{\Delta}_o). \tag{4.13}$$

For BART+SPL, Nethery et al. (2019) also introduced an unidentifiable variance inflation parameter for observations in the region of non-overlap by adding a variance component based on the propensity score to σ^2_{Smo}. More precisely, they use

$$\epsilon_o \sim \mathcal{N}\left(0, \sigma^2_{Smo} + \mathbb{1}(\widehat{p.sc}_o \in O^\neg) \cdot \text{var}_{\text{infl}}\right) \tag{4.14}$$

for the noise term in the smoothing model in Equation (4.6). The idea of this proposal is to satisfy the need of an increase in uncertainty where overlap is weak due to sparse data and the fact that BART itself does not properly account for this need as pointed out in Section 3.3. The BART algorithm does not itself indicate the higher uncertainty when there is weak overlap such that var_{infl} is proposed as a remedy for this issue in Nethery et al. (2019). However, the parameter is unidentifiable and treated as a hyperparameter with rough guidelines on how to specify default values for it. In contrast to that, Appendix B suggests that the SoftBART algorithm is capable to reflect the higher uncertainty in regions of weak overlap. Therefore, the SBART+SPL algorithm gets rid of this unidentifiable variance inflation parameter in step 8.

Step 9 finishes one iteration of the MCMC algorithm by drawing the average treatment effect estimate $\hat{\Delta}^{(m)}_P$ via Bayesian Bootstrap as outlined in Wang et al. (2015) to account for uncertainty in confounder and effect modifier selection. For each $\hat{\Delta}^{(m)}_P$, the set of individual treatment effects $\left\{\tilde{\Delta}^{(m)}_O, \tilde{\Delta}^{(m)}_{O^\neg}\right\}$ is bootstrapped $B = 250$ times and the B averages of these bootstrap draws are draws from the posterior distribution of the population average treatment effect. A random sample of these B averages is then termed $\hat{\Delta}^{(m)}_P$. The Bayesian Bootstrap is motivated for treatment effect estimation in settings with many potential confounders but small sample size, probable interactions between confounders and the

treatment variable as well as uncertainty on the inclusion of the particular confounders and interaction terms. The naive approach of calculating an average treatment effect would be to take a simple average over the draws of the individual treatment effects $\left\{\tilde{\Delta}_{O}^{(m)}, \tilde{\Delta}_{O^{\neg}}^{(m)}\right\}$. This places uniform mass on each of the confounder vectors. The Bayesian Bootstrap procedure accounts for variability in the empirical estimate as the weight of each confounder stratum is considered as an unknown parameter with own prior distribution (Wang et al. 2015).

5 Simulation Study

This simulation study investigates the behavior of SBART+SPL compared to BART+SPL for inference regarding τ_P with high-dimensional covariates as already presented similarly in the simulation study of Nethery et al. (2019). As highlighted in Section 3, a key property of the underlying SoftBART algorithm as an integral part of SBART+SPL is variable selection in sparse datasets which may give SBART+SPL an edge over BART+SPL in terms of precision concerning inference on τ_P. Added to that, Nethery et al. (2019) indicate severe undercoverage and increasing bias for BART+SPL when the number of covariates rises. SBART+SPL is expected to improve credible interval coverage compared to BART+SPL as elaborated in Section 4.2.

The target estimand τ_P is computed by its corresponding population average treatment effect estimate, $\hat{\Delta}_P$, as outlined in Section 4. The simulation study follows the notation of Zhu et al. (2022) such that

$$\hat{\Delta}_P^{(sim)} = \frac{1}{M} \sum_{m=1}^{M} \hat{\Delta}_P^{(m)}, \tag{5.1}$$

where $\hat{\Delta}_P^{(sim)}$ is an estimate of the population average treatment effect for the specific dataset $sim \in \{1, ..., Sim\}$ with $Sim = 200$. As a default, $M = 5,000$ posterior draws

(after $10,000$ burn-in draws) are chosen for BART+SPL and SBART+SPL. Based on the true two potential outcomes for each observation i, one can calculate the true individual treatment effect $Y_i(1) - Y_i(0)$ for each individual i. The true average treatment effect for each simulation dataset reads $ace_{true}^{(sim)} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$. Consequently, we can evaluate the absolute bias and mean-squared error over all simulations:

$$|\text{Bias}| = \frac{1}{Sim}\sum_{sim=1}^{Sim}\left|\hat{\Delta}_P^{(sim)} - ace_{true}^{(sim)}\right|, \tag{5.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{Sim}\sum_{sim=1}^{Sim}\left(\hat{\Delta}_P^{(sim)} - ace_{true}^{(sim)}\right)^2}. \tag{5.3}$$

Besides precision, coverage rates evaluate the efficiency of the respective estimation methods. For each simulation dataset, the indicator function checks if the true average treatment effect lies in $CI_{95}^{(sim)}$, defined as the 95% credible interval based on the posterior draws $\hat{\Delta}_P^{(m)}$. Moreover, the average width of the 95% credible interval indicates the tightness of the intervals.

$$\text{Coverage}_{95\%-CI} = \frac{1}{Sim}\sum_{sim=1}^{Sim}\mathbb{1}\left(ace_{true}^{(sim)} \in CI_{95}^{(sim)}\right), \tag{5.4}$$

$$\text{Width}_{95\%-CI} = \frac{1}{Sim}\sum_{sim=1}^{Sim}\text{length}\left(CI_{95}^{(sim)}\right). \tag{5.5}$$

The data generating process of this simulation study follows the simulation study concerning high-dimensional covariates of Nethery et al. (2019). There are ten true confounding variables $X_{1i}, ..., X_{10i}$ which follow the subsequent Bernoulli and Normal distributions dependent on the treatment status D_i:

$$X_{1i}|D_i = 1, ..., X_{5i}|D_i = 1 \sim \text{Bernoulli}(0.45),$$

$$X_{1i}|D_i = 0, ..., X_{5i}|D_i = 0 \sim \text{Bernoulli}(0.4),$$

$$X_{6i}|D_i = 1, ..., X_{10i}|D_i = 1 \sim \mathcal{N}(2, 4),$$

$$X_{6i}|D_i = 0, ..., X_{10i}|D_i = 0 \sim \mathcal{N}(1.3, 1).$$

Based on the true confounder variables from above, we produce the two potential outcomes for each individual $i = 1, ..., n$ with $n = 500$:

$$\begin{aligned} Y_i(0) &= 0.5\left(X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i}\right) + \\ &\quad \frac{15}{(1 + \exp\{-8X_{6i} + 1\})} + X_{7i} + X_{8i} + X_{9i} + X_{10i} - 5, \\ Y_i(1) &= X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i} - \\ &\quad 0.5\left(X_{6i} + X_{7i} + X_{8i} + X_{9i} + X_{10i}\right). \end{aligned}$$

Added to the 10 true confounding variables from above, a set of covariates with size $ncov \in \{10, 25, 50\}$ are included into the dataset and drawn independently from a standard Normal distribution irrespective the treatment status D_i. Figures B.1, B.2, B.3 in Appendix B illustrate the corresponding propensity score estimates for selected simulation datasets. The region of overlap and non-overlap is defined by using the approach of Nethery et al. (2019) that is described in Subsection 2.2. The figures show that there is a substantial region of non-overlap such that BART+SPL and SBART+SPL have to extrapolate into these regions.

Besides BART+SPL, SBART+SPL is compared to three further benchmark methods as described similarly in Nethery et al. (2019). For each benchmark method, there is an untrimmed and a trimmed version implemented[4] such that 8 methods are compared in

[4]The trimmed versions just use observations within the region of overlap to estimate PATE. If we

total:

- **U-GR, T-GR**: The GR approach implements the method of multiple imputation with two subclassification splines (MITSS) proposed by Gutman & Rubin (2015).
- **U-BART, T-BART**: This fits a single BART model with covariates, exposure variable and estimated propensity scores to the outcome variable. Afterwards, potential outcomes are predicted based on the posterior predictive distributions and then used to compute the treatment effect (Nethery et al. 2019).
- **U-SoftBART, T-SoftBART**: The procedure is similar to the previous BART approach but uses SoftBART (Linero 2022) instead of BART for modeling.
- **BART+SPL**: The original approach proposed by Nethery et al. (2019).
- **SBART+SPL**: The SBART+SPL approach proposed in this paper and mainly described in Section 4.

The precision and coverage values are shown in Table 1 for 10, 25, and 50 additional covariates. The trimmed and untrimmed benchmark models that use BART and SBART for modeling seem to be heavily biased across settings. Moreover, the credible intervals of these methods seem to be unable to cover the true PATE due to the biased estimates and the tight credible intervals. The U-GR approach gives low values for RMSE and absolute bias in relation to competitors and appears to improve with an increase in the number of additional covariates from 10 to 50. However, especially for $ncov = 10$ and $ncov = 25$, U-GR deviates decisively from the nominal coverage rate of 0.95. BART+SPL shows precision values that are comparable to the results of T-GR but has coverage properties better than T-GR and even U-GR, although still not even close to the nominal coverage rate. The

use a trimmed version of a method, this method is abbreviated **T-method** and its untrimmed version is abbreviated **U-method**.

improved coverage of BART+SPL may be partly explained by larger credible intervals, which are roughly twice as long as the intervals of U-GR.

SBART+SPL improves upon BART+SPL in all measured dimensions and for all numbers of additional covariates. The method shows lower values for RMSE and absolute bias while just U-GR outperforms SBART+SPL for $ncov = 25$ and $ncov = 50$ in this regard. The coverage values are the closest to the nominal level of 0.95 across all competitor models and across all settings, although the width of the credible intervals is smaller. However, the issue of undercoverage is still present. Moreover, SBART+SPL remains relatively robust in terms of precision and coverage with regard to an increase in the number of additional covariates.

6 Empirical Application: NG compressor stations, cancer and mortality in the US

This section revisits the empirical application in Nethery et al. (2019) where the effect of Natural Gas (NG) compressor stations on cancer and mortality is analyzed. More detailed information about data collection can be found in their publication and the original description of Mokdad et al. (2017). The research question is whether the existence of NG compressor stations in United States counties affects the county-level mortality rates for thyroid cancer and leukemia. The study focuses on the mid-western region in the US to exclude confounding alternative pollutions that are argued to be more prevalent in the coastal area. Moreover, exposure to pollution by NG compressor stations might be more relevant in the mid-west due to longer industrial history. To further unconfound the estimation, many pre-treatment variables ($p = 22$) are considered that represent demographic, socio-economic and behavioral characteristics of each county.

Summaries of the four outcome variables of mortality rates and the explanatory variables

are presented in Table 5 and Table 6 in Appendix C. In total, $n = 978$ counties with full confounder information are considered. The treatment group consists of 291 counties where a county is considered as being exposed to treatment when at least one NG compressor station exists in this county. On the contrary, 687 counties are considered to be unexposed. The research question rests on the assumption that the minimum latency period of thyroid cancer (2.5 years) and leukemia (0.4 years) is covered by the analysis. It can be verified that most of the NG compressor stations have their highest operating dates before or in 2012. That is, the mortality rates of thyroid cancer and leukemia observed in 2014 are argued to cover their minimum latency periods regarding the exposure to emissions from NG compressor stations. The histogram of propensity score estimates in Figure C.1 in Appendix C indicates that the region of non-overlap cannot be neglected as roughly 13% of the observations fall into that region when defined by the approach of Nethery et al. (2019) described in Subsection 2.2. Tables 2 and 3 show PATE estimates and its corresponding credible intervals where SBART+SPL is compared to BART+SPL and the benchmark methods described in the simulation study of Section 5. Note that for the trimmed methods (T-GR, T-BART, T-SoftBART) the target estimand changes from a population to a sample average treatment effect that omits the observations in the region of non-overlap. The considered outcome variables in Table 2 are (1) log-transformed 2014 leukemia mortality rate and (2) percent point change in leukemia mortality rate from 1980 to 2014 whereas Table 3 considers the (3) log-transformed 2014 thyroid cancer mortality rate and (4) percent point change in thyroid cancer mortality rate from 1980 to 2014.

Across both tables, one observes for SBART+SPL positive treatment effects where the 95% credible interval excludes the null in three out of four cases (no significant positive effect for (3), the log-transformed 2014 thyroid cancer mortality rate). All other methods report no significant treatment effect except for untrimmed and trimmed BART regarding the change in thyroid cancer mortality rates. BART+SPL seems to inflate the width

of the credible intervals, compared to BART or SoftBART, due to the variance inflation parameter within its algorithm and the increase in uncertainty by extrapolating into the non-overlap region. A positive effect must be strong when faced with these enlarged intervals and seems to be insufficient for BART+SPL to retrieve a positive treatment effect with corresponding credible intervals that exclude the null. However, the interval width might still be too small for BART+SPL as the simulation study in Section 5 detects undercoverage issues even if the number of covariates equals ten and not $p = 22$ as in this empirical study. For SBART+SPL, the credible intervals are even larger compared to the intervals of BART+SPL. However, in three out of four cases, SBART+SPL seems to find enough compelling evidence in the data to support a positive treatment effect of exposure to natural gas compressor stations to an increase in mortality rates, although the uncertainty around the exact point estimate is quite large. In general, almost all competitor methods report credible intervals including the null but positive treatment effects of exposure to NG compressor stations on the mortality rates of leukemia and thyroid cancer. At least, this justifies a more detailed investigation of these health effects with finer spatial data. Moreover, other data-related problems like the potentially close connection of NG compressor station locations and natural gas drilling station locations as well as the more insightful (but not available) information about cancer diagnosis instead of mortality rates might affect and influence the analysis presented here and any policy-related measures derived from it (Nethery et al. 2019).

7 Conclusion

Violations of the positivity assumption in the potential outcome framework affect the estimation of the population average treatment effect. While remedies like trimming and weighting often change the initial target estimand, extrapolation approaches might be useful

if a researcher wants to do inference with the original observational data at hand. This paper proposes a new method, SBART+SPL, which builds on the framework of BART+SPL by Nethery et al. (2019) but adapts more appropriately to an increasing number of covariates. In Section 4, the SBART+SPL algorithm is described as an MCMC procedure that estimates the population average treatment effect by discriminating between a region of overlap and non-overlap. The contrast to BART+SPL is mainly twofold. First, instead of using BART for the imputation stage, the SoftBART algorithm introduced by Linero & Yang (2018) computes individual causal effects in the region of overlap. Second, the spline model in the extrapolation phase does not rely on an unidentifiable variance inflation parameter to properly account for higher uncertainty in regions of non-overlap, as it is done in the BART+SPL algorithm.

A first simulation study in Section 5, that illustrates the violation of the positivity assumption and an increase in the number of covariates, suggests improvements in precision and coverage compared to BART+SPL and benchmark methods, even though nominal coverage is still not reached fully. An extension to use SBART+SPL with binary outcomes is provided in the Appendix D. Section 6 demonstrates the applicability of SBART+SPL for continuous outcomes by estimating the population average treatment effect of NG compressor stations on leukemia and thyroid cancer mortality rates on US county level. Unlike BART+SPL and baseline methods, SBART+SPL appears to be able to recover a positive treatment effect from data with, however, rather large uncertainty around point estimates compared to the competitor methods.

References

Austin, P. C. (2011), 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies', *Multivariate Behavioral Research*

46(3), 399–424.

URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/*

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

URL: *https://doi.org/10.1007/BF00058655*

Breiman, L. (2001), 'Random Forests', *Machine Learning* **45**(1), 5–32.

URL: *https://doi.org/10.1023/A:1010933404324*

Chipman, H. A., George, E. I. & McCulloch, R. E. (1998), 'Bayesian CART Model Search', *Journal of the American Statistical Association* **93**(443), 935–948. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

URL: *https://www.jstor.org/stable/2669832*

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010), 'BART: Bayesian additive regression trees', *The Annals of Applied Statistics* **4**(1), 266–298. Publisher: Institute of Mathematical Statistics.

URL: *https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-1/BART-Bayesian-additive-regression-trees/10.1214/09-AOAS285.full*

Chipman, H., George, E. & Mcculloch, R. (2006), Bayesian Ensemble Learning, *in* 'Advances in Neural Information Processing Systems', Vol. 19, MIT Press.

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**(1), 187–199.

URL: *https://doi.org/10.1093/biomet/asn055*

D'Amour, A., Ding, P., Feller, A., Lei, L. & Sekhon, J. (2020), 'Overlap in Observational Studies with High-Dimensional Covariates', *arXiv:1711.02582 [math, stat]* . arXiv: 1711.02582.

URL: *http://arxiv.org/abs/1711.02582*

Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. (2019), 'Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition', *Statistical Science* **34**(1), 43–68. Publisher: Institute of Mathematical Statistics.
URL: *https://www.jstor.org/stable/26771031*

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine.', *The Annals of Statistics* **29**(5), 1189–1232. Publisher: Institute of Mathematical Statistics.
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full*

Gutman, R. & Rubin, D. (2013), 'Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes', *Statistics in Medicine* **32**(11), 1795–1814. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5627.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5627*

Gutman, R. & Rubin, D. B. (2015), 'Estimation of causal effects of binary treatments in unconfounded studies', *Statistics in medicine* **34**(26), 3381–3398.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782596/*

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2020), 'Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)', *Bayesian Analysis* **15**(3), 965–1056. Publisher: International Society for Bayesian Analysis.
URL: *https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference–Regularization-Confounding/10.1214/19-BA1195.full*

Hernán, M. A. & Robins, J. M. (2020), *Causal Inference: What If*, 1st edn, Boca Raton: Chapman & Hall/CRC.

Hernán, M. A. & Robins, J. M. (2006), 'Estimating causal effects from epidemiological data', *Journal of Epidemiology and Community Health* **60**(7), 578–586.

Hill, J., Linero, A. & Murray, J. (2020), 'Bayesian Additive Regression Trees: A Review and Look Forward', *Annual Review of Statistics and Its Application* **7**(1), 251–278.
URL: *https://www.annualreviews.org/doi/10.1146/annurev-statistics-031219-041110*

Hill, J. & Su, Y.-S. (2013), 'Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes', *The Annals of Applied Statistics* **7**(3). arXiv: 1311.7244.
URL: *http://arxiv.org/abs/1311.7244*

Hu, L., Liu, B., Ji, J. & Li, Y. (2020), 'Tree-Based Machine Learning to Identify and Understand Major Determinants for Stroke at the Neighborhood Level', *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease* **9**(22), e016745.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7763737/*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021), *An introduction to statistical learning : with applications in R*, Springer texts in statistics, second edition edn, Springer New York, NY, New York, NY. Section: xv, 607 pages : illustrations (chiefly color) ; 24 cm.

Krantsevich, N., He, J. & Hahn, P. R. (2023), Stochastic tree ensembles for estimating heterogeneous effects, *in* F. Ruiz, J. Dy & J.-W. van de Meent, eds, 'Proceedings of The 26th International Conference on Artificial Intelligence and Statistics', Vol. 206 of *Proceedings of Machine Learning Research*, PMLR, pp. 6120–6131.
URL: *https://proceedings.mlr.press/v206/krantsevich23a.html*

Li, F., Ding, P. & Mealli, F. (2022), 'Bayesian Causal Inference: A Critical Review'. arXiv:2206.15460 [stat].

URL: *http://arxiv.org/abs/2206.15460*

Li, F., Morgan, K. L. & Zaslavsky, A. M. (2018), 'Balancing Covariates via Propensity Score Weighting', *Journal of the American Statistical Association* **113**(521), 390–400. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2016.1260466.

URL: *https://doi.org/10.1080/01621459.2016.1260466*

Li, F., Thomas, L. E. & Li, F. (2019), 'Addressing Extreme Propensity Scores via the Overlap Weights', *American Journal of Epidemiology* **188**(1), 250–257.

URL: *https://doi.org/10.1093/aje/kwy201*

Linero, A. R. (2018), 'Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection', *Journal of the American Statistical Association* **113**(522), 626–636. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1264957.

URL: *https://doi.org/10.1080/01621459.2016.1264957*

Linero, A. R. (2022), 'SoftBart: Soft Bayesian Additive Regression Trees'. arXiv:2210.16375 [stat].

URL: *http://arxiv.org/abs/2210.16375*

Linero, A. R. & Yang, Y. (2018), 'Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80**(5), 1087–1110.

URL: *https://academic.oup.com/jrsssb/article/80/5/1087/7048381*

Mokdad, A. H., Dwyer-Lindgren, L., Fitzmaurice, C., Stubbs, R. W., Bertozzi-Villa, A., Morozoff, C., Charara, R., Allen, C., Naghavi, M. & Murray, C. J. L. (2017), 'Trends and Patterns of Disparities in Cancer Mortality Among US Counties, 1980-2014', *JAMA*

317(4), 388–406.

URL: *https://doi.org/10.1001/jama.2016.20324*

Neal, R. M. (2003), 'Slice sampling', *The Annals of Statistics* **31**(3), 705–767. Publisher: Institute of Mathematical Statistics.

URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-31/issue-3/Slice-sampling/10.1214/aos/1056562461.full*

Nethery, R. C., Mealli, F. & Dominici, F. (2019), 'ESTIMATING POPULATION AVERAGE CAUSAL EFFECTS IN THE PRESENCE OF NON-OVERLAP: THE EFFECT OF NATURAL GAS COMPRESSOR STATION EXPOSURE ON CANCER MORTALITY', *The annals of applied statistics* **13**(2), 1242–1267.

URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6658123/*

Pratola, M. T. (2016), 'Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian Regression Tree Models', *Bayesian Analysis* **11**(3), 885–911. Publisher: International Society for Bayesian Analysis.

URL: *https://projecteuclid.org/journals/bayesian-analysis/volume-11/issue-3/Efficient-MetropolisHastings-Proposal-Mechanisms-for-Bayesian-Regression-Tree-Models/10.1214/16-BA999.full*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

URL: *https://www.R-project.org/*

Rockova, V. & van der Pas, S. (2019), 'Posterior Concentration for Bayesian Regression Trees and Forests'. arXiv:1708.08734 [math, stat].

URL: *http://arxiv.org/abs/1708.08734*

Rubin, D. B. (2005), 'Causal Inference Using Potential Outcomes: Design, Modeling, Decisions', *Journal of the American Statistical Association* **100**(469), 322–331. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
URL: *https://www.jstor.org/stable/27590541*

Saha, E. (2023), 'Theory of Posterior Concentration for Generalized Bayesian Additive Regression Trees'. arXiv:2304.12505 [math].
URL: *http://arxiv.org/abs/2304.12505*

Stürmer, T., Rothman, K. J., Avorn, J. & Glynn, R. J. (2010), 'Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study', *American Journal of Epidemiology* **172**(7), 843–854.

Visconti, G. & Zubizarreta, J. R. (2018), 'Handling Limited Overlap in Observational Studies with Cardinality Matching', *Observational Studies* **4**(1), 217–249. Publisher: University of Pennsylvania Press.
URL: *https://muse.jhu.edu/pub/56/article/793377*

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.
URL: *https://books.google.de/books?id=ScRQJEETs0EC*

Wang, C., Dominici, F., Parmigiani, G. & Zigler, C. M. (2015), 'Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models', *Biometrics* **71**(3), 654–665.

Wang, M., He, J. & Hahn, P. R. (2024), 'Local gaussian process extrapolation for bart models with applications to causal inference', *Journal of Computational and Graphical*

Statistics **33**(2), 724–735.

URL: *https://doi.org/10.1080/10618600.2023.2240384*

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H. & Gallego, B. (2018), 'Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases', *Statistics in Medicine* **37**(23), 3309–3324. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7820.

URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7820*

Westreich, D. & Cole, S. R. (2010), 'Invited Commentary: Positivity in Practice', *American Journal of Epidemiology* **171**(6), 674–677.

URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2877454/*

Wu, Y., , Håkon, T., & West, M. (2007), 'Bayesian CART: Prior Specification and Posterior Simulation', *Journal of Computational and Graphical Statistics* **16**(1), 44–66. Publisher: ASA Website _eprint: https://doi.org/10.1198/106186007X180426.

URL: *https://doi.org/10.1198/106186007X180426*

Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J. & Mitra, N. (2021), 'Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches', *Pharmacoepidemiology and drug safety* **30**(11), 1471–1485.

URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8492528/*

Zhu, Y., Mitra, N. & Roy, J. (2022), 'Addressing Positivity Violations in Causal Effect Estimation using Gaussian Process Priors'. arXiv:2110.10266 [stat].

URL: *http://arxiv.org/abs/2110.10266*

A Properties of SoftBART

A.1 Hard and soft decision rules

The contrast between hard decision rules and soft decision rules as in Equation (3.4) and Equation (3.8) can be visualized in a simple prediction example. Let the univariate regressor be a sequence of values ranging from 0 to 1, incremented by 0.01, such that $X_i = (0, 0.01, 0.02, ..., 1)$. The continuous outcome variable is generated with

$$Y_i(X_i = x) = \sin(2\pi x) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 0.1^2). \tag{A.1}$$

In-sample predictions of the outcome variable using BART and SoftBART are given in Figure A.1. Whereas the BART predictions wiggle non-smoothly around the true sine curve, the SoftBART predictions result in a similarly shaped curve compared to the true sine curve. This results in a lower RMSE value for the SoftBART algorithm (0.046) compared to the BART algorithm (0.067).

The improvement in RMSE is not an artifact of the underlying smooth sine curve that generates our outcome. Let us change our outcome model to a step function by:

$$Y_i(X_i = x) = 2 - 4 \cdot \mathbb{1}(x > 0.5) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 0.1^2). \tag{A.2}$$

In-sample predictions of the outcome variable using BART and SoftBART are given in figure A.2. The RMSE for SoftBART (1.360) is slightly lower compared to the RMSE of BART (1.366), although the outcome model mimics a step function with a hard jump from 2 to -2 around $x = 0.5$.

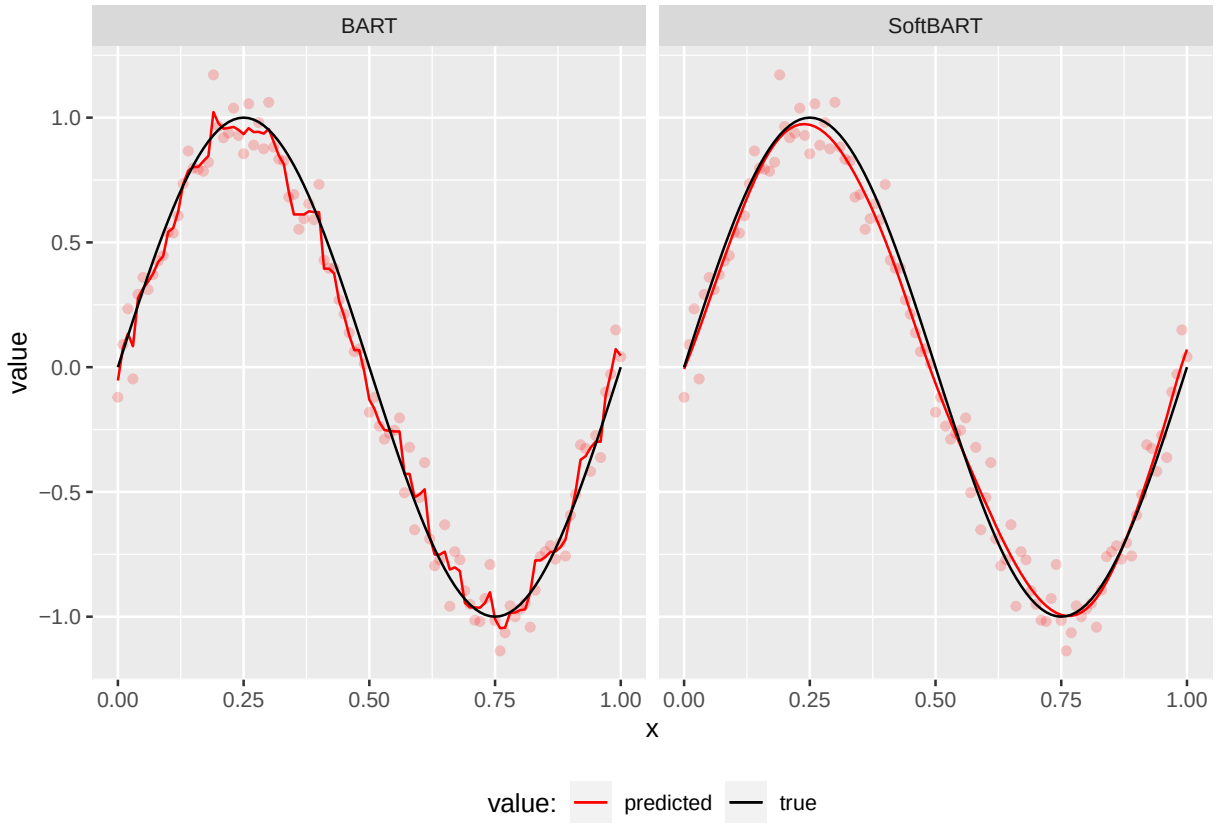

Figure A.1: In-sample predictions for BART and SoftBART. Outcomes are generated by a sine curve.

A.2 Uncertainty quantification

Added to improved precision, the SoftBART algorithm seems to be able to improve uncertainty quantification when estimating treatment effects. This section exemplifies this improvement by recollecting a simple example of uncertainty quantification around conditional average treatment effects and can be found similarly in the discussion part of Hahn et al. (2020). Let us generate $n = 200$ observations, 100 units for each treatment group $D_i \in \{0, 1\}$. The univariate regressor $X_i \sim \text{Ga}(\mu_{Ga}, sd = 8)$ has $\mu_{Ga} = 35$ for the treated units and $\mu_{Ga} = 60$ for the control units. Based on the regressor, a linear outcome model is used to generate the respective outcome values by

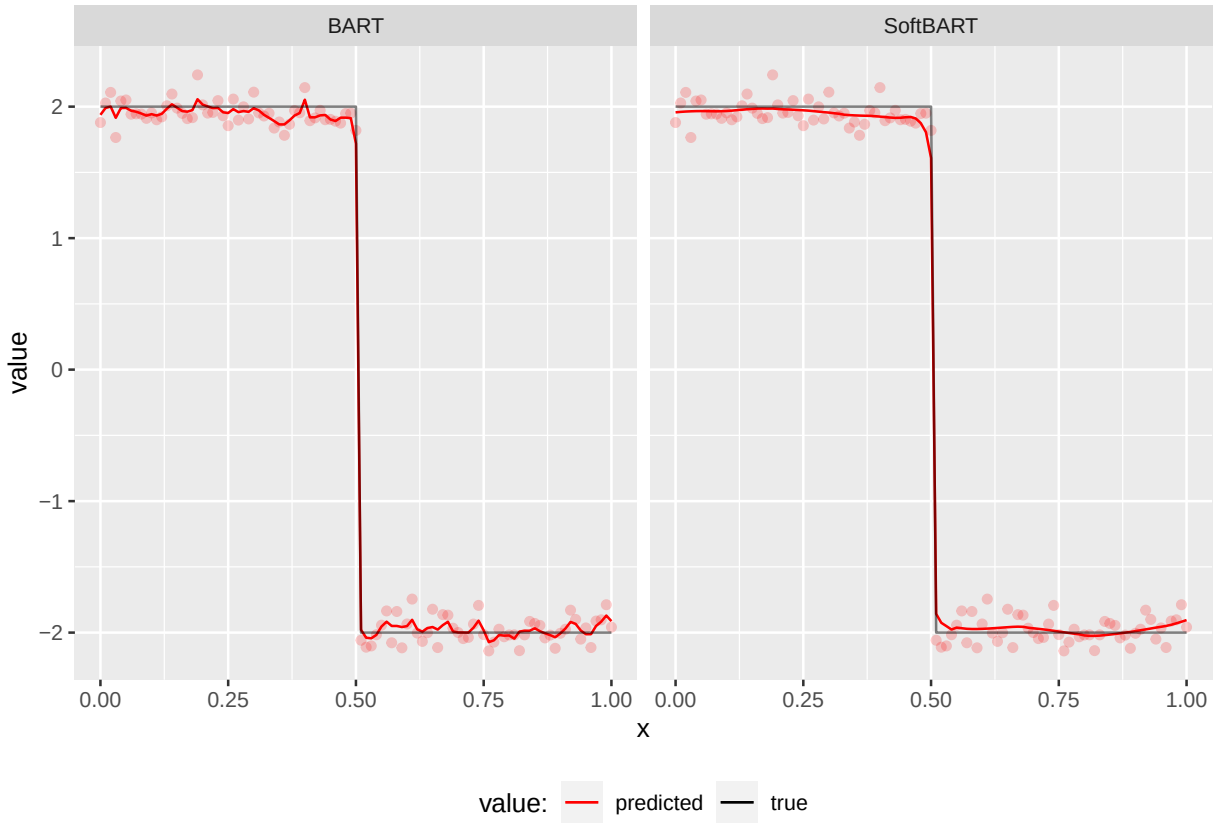

Figure A.2: In-sample predictions for BART and SoftBART. Outcomes are generated by a step function.

$$Y_i(X_i = x, D_i = d) = 10 + 5 \cdot d + 0.3 \cdot x + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1). \tag{A.3}$$

Figure A.3 displays observed and predicted outcome values for each treatment group separately and estimates of the population conditional average causal effects $\tau_{P|\mathbf{X}}$ for the SoftBART as well as for the BART algorithm. The observed values are overlapping for both treatment groups for values lying in the interval $X_i \approx (40, 60)$ but non-overlapping for values outside of this interval. Consequently, estimation methods should quantify this uncertainty in the non-overlapping region with increasing credible intervals around $\hat{\tau}_{P|\mathbf{X}}$. Whereas SoftBART accounts for that feature, BART does not transmit the increased uncertainty into its credible intervals resulting in too overconfident predictions and treatment

effect estimates.

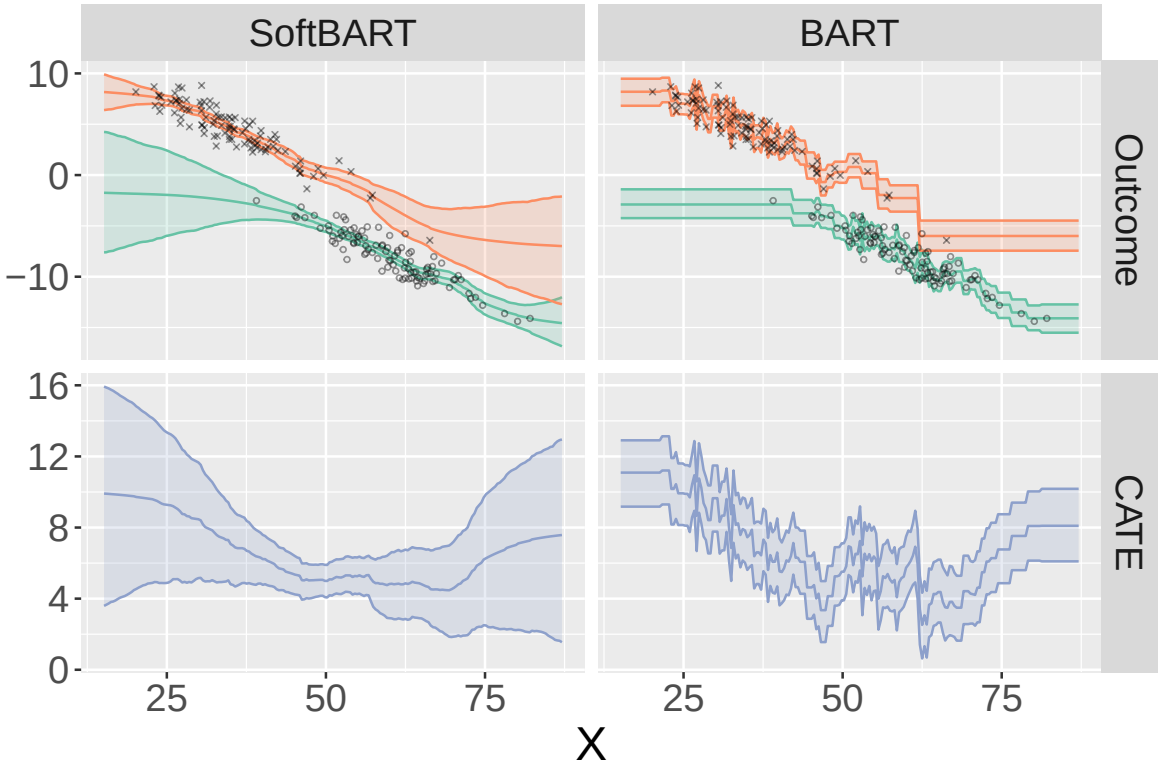

Figure A.3: Comparison of uncertainty quantification for BART and SoftBART.

B Simulation Study: Descriptive Statistics

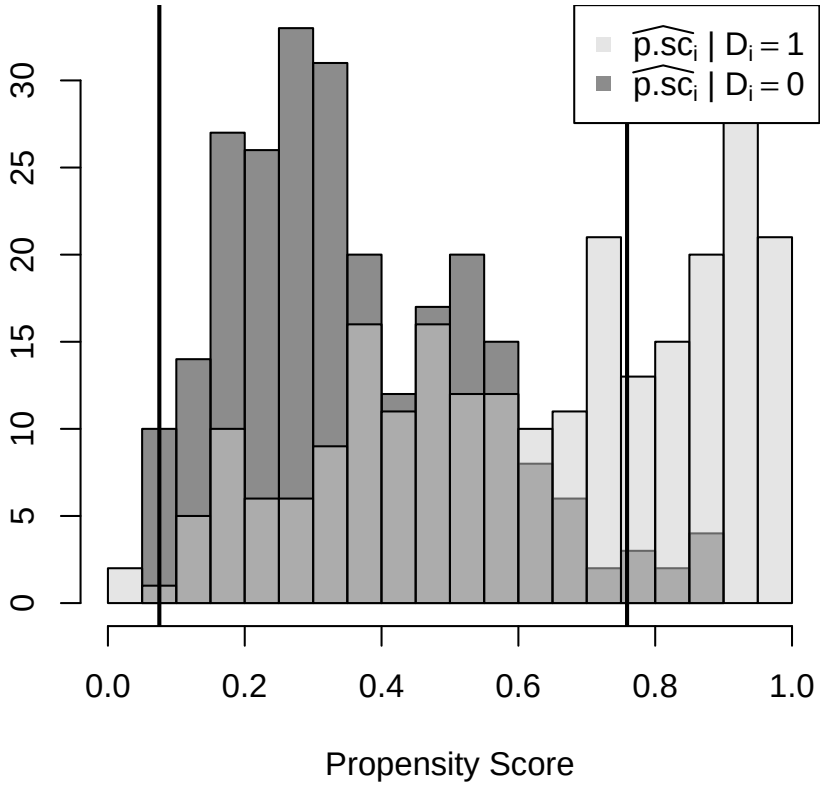

Figure B.1: Example of a distribution of propensity score estimates by treatment status for $ncov = 10$. Regions of overlap and non-overlap are indicated by the vertical straight lines. The region of overlap lies in-between these two vertical straight lines.

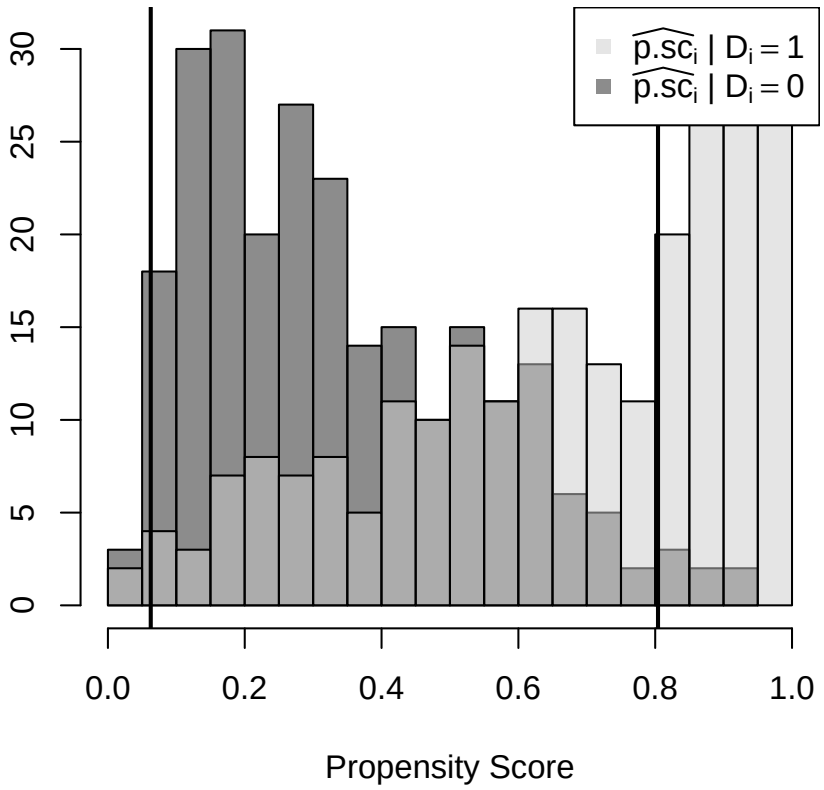

Figure B.2: Example of a distribution of propensity score estimates by treatment status for $ncov = 25$. The region of overlap lies in-between these two vertical straight lines.

Table 1: Precision and coverage values for 200 simulated datasets with sample size $n = 500$ for each dataset.

ncov	Method	RMSE	|Bias|	$\text{Coverage}_{95\%-CI}$	$\text{Width}_{95\%-CI}$
10	U-GR	0.8092	0.7455	0.21	0.9698
	T-GR	1.1657	1.1075	0.00	0.6370
	U-BART	17.4367	17.4340	0.00	0.1526
	T-BART	17.4362	17.4335	0.00	0.1715
	U-SoftBART	17.4364	17.4337	0.00	0.2880
	T-SoftBART	17.4361	17.4334	0.00	0.3012
	BART+SPL	1.5036	1.2132	0.50	2.2323
	SBART+SPL	0.7510	0.6477	0.72	1.9032
25	U-GR	0.5389	0.4559	0.55	0.9763
	T-GR	1.1691	1.1193	0.00	0.7033
	U-BART	17.4608	17.4608	0.00	0.1940
	T-BART	17.4592	17.4592	0.00	0.2212
	U-SoftBART	17.4619	17.4595	0.00	0.3311
	T-SoftBART	17.4610	17.4586	0.00	0.3530
	BART+SPL	1.6585	1.2466	0.45	2.0152
	SBART+SPL	0.7313	0.6463	0.83	1.7782
50	U-GR	0.4234	0.3447	0.71	1.0493
	T-GR	1.1598	1.0686	0.00	0.7657
	U-BART	17.3484	17.3468	0.00	0.2531
	T-BART	17.3485	17.3469	0.00	0.2878
	U-SoftBART	17.3514	17.3498	0.00	0.3364
	T-SoftBART	17.3500	17.3483	0.00	0.4215
	BART+SPL	1.2761	1.0223	0.64	1.9134
	SBART+SPL	0.8031	0.6258	0.79	1.7706

Table 2: Average treatment effect estimates and corresponding 95% credible intervals for outcome variables (1) log-transformed 2014 leukemia mortality rate, and (2) percent point change in leukemia mortality rate from 1980 to 2014.

	Method	Effect	95%-CI		95%-CI-Width
(1)	U-GR	0.009	-0.004	0.021	0.025
	T-GR	0.005	-0.004	0.014	0.019
	U-BART	0.004	-0.006	0.013	0.019
	T-BART	0.002	-0.004	0.008	0.012
	U-SoftBART	0.001	-0.004	0.006	0.010
	T-SoftBART	0.009	-0.002	0.020	0.022
	BART+SPL	0.001	-0.024	0.025	0.049
	SBART+ SPL	0.198	0.023	1.003	0.981
(2)	U-GR	1.729	0.569	3.058	2.489
	T-GR	0.935	-0.024	1.890	1.914
	U-BART	0.915	-0.051	1.889	1.940
	T-BART	0.634	-0.147	1.617	1.764
	U-SoftBART	0.577	-0.156	1.637	1.793
	T-SoftBART	1.613	0.415	2.769	2.353
	BART+SPL	0.508	-1.576	2.519	4.095
	SBART+ SPL	15.449	1.862	28.610	26.748

Table 3: Average treatment effect estimates and corresponding 95% credible intervals for outcome variables (3) log-transformed 2014 thyroid cancer mortality rate, and (4) percent point change in thyroid cancer mortality rate from 1980 to 2014.

	Method	Effect	95%-CI		95%-CI-Width
(3)	U-GR	-0.014	-0.024	-0.003	0.021
	T-GR	0.003	-0.006	0.012	0.018
	U-BART	0.003	-0.006	0.012	0.018
	T-BART	-0.001	-0.008	0.006	0.013
	U-SoftBART	-0.001	-0.007	0.005	0.012
	T-SoftBART	-0.013	-0.024	-0.003	0.021
	BART+SPL	-0.005	-0.029	0.019	0.047
	SBART+SPL	0.104	-1.024	1.039	2.063
(4)	U-GR	0.549	-0.610	1.730	2.340
	T-GR	1.053	0.111	1.997	1.887
	U-BART	1.047	0.088	1.986	1.898
	T-BART	0.806	-0.032	1.820	1.852
	U-SoftBART	0.747	-0.082	1.843	1.924
	T-SoftBART	0.558	-0.416	1.509	1.925
	BART+SPL	0.824	-1.121	2.726	3.847
	SBART+SPL	15.499	2.441	28.208	25.767

Method	Effect	95%-CI		95%-CI-Width
U-GR	-0.016	-0.083	0.050	0.133
U-BART	-0.002	-0.037	0.017	0.054
U-SoftBART	-0.013	-0.086	0.058	0.144
T-BART	-0.002	-0.049	0.025	0.075
T-SoftBART	0.012	-0.052	0.069	0.122
SBCF	-0.009	-0.075	0.048	0.123
BCF	-0.010	-0.080	0.048	0.129
BART+SPL	-0.008	-0.065	0.039	0.105
SBART+ SPL	0.033	0.000	0.176	0.176

Table 4: Average treatmet effect of right heart catheterization (RHC) within the first 24 hours of study entry on the 180-day survival of critically ill female patients.

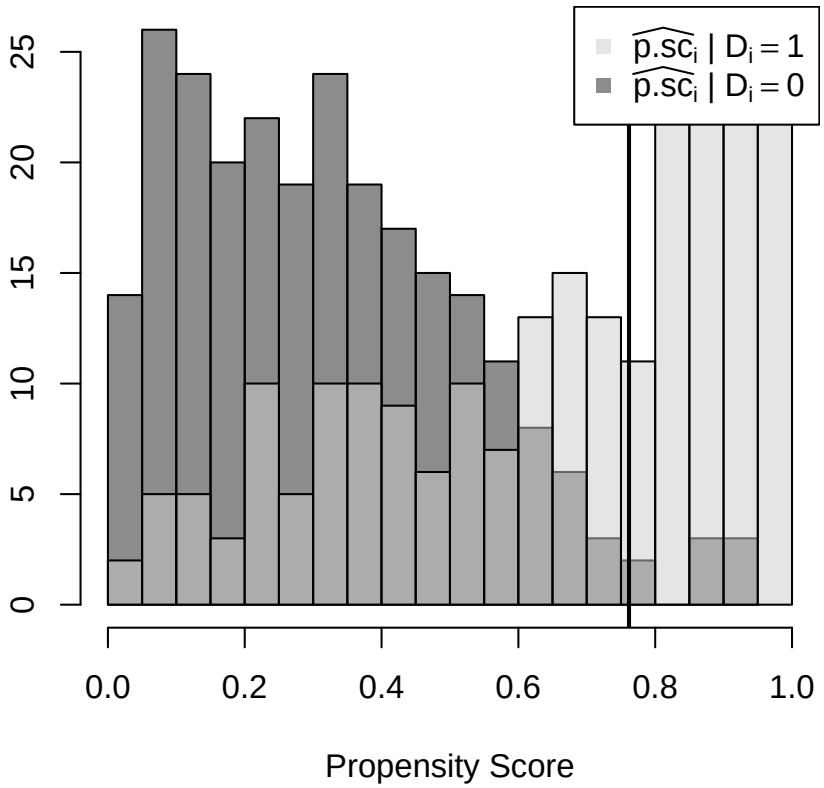

Figure B.3: Example of a distribution of propensity score estimates by treatment status for $ncov = 25$. The region of overlap lies to the left of the vertical straight line as there are just extreme propensity values close to 1 in this example.

C Empirical Application: Descriptive Statistics

Table 5: County-level ($n = 978$) descriptive statistics of the four outcome variables. 2014 mortality rates are measured as deaths per $100,000$ population, the changes in mortality rates are measured as percentage points.

Variable	Mean	St. Dev.	Min	Max
leukemia mortality rate				
2014 (log-transformed)	9.56 (2.25)	0.10 (0.11)	4.17 (1.43)	16.55 (2.81)
change from 1980 to 2014	−2.02	8.98	−42.94	40.14
thyroid cancer mortality rate				
2014 (log-transformed)	0.56 (−0.60)	0.06 (0.09)	0.30 (−0.89)	0.930 (−0.07)
change from 1980 to 2014	1.71	8.03	−20.11	29.06

Table 6: County-level ($n = 978$) descriptive statistics of covariate variables ($p = 22$).

Variable	Mean	St. Dev.	Min	Max
population per prim. care physician	801.39	1,330.09	60	13,664
less than 65 year olds uninsured (%)	17.74	5.53	4.72	38.85
diabetic (%)	83.02	8.14	25.46	100.00
current smokers (%)	20.51	6.17	6.60	49.20
limited access to healthy foods (%)	10.18	8.19	0.00	62.94
obese (%)	28.43	5.19	10.40	43.50
food environment index	7.31	1.32	0.61	9.84
population per mi^2	99.38	332.66	0.52	5,144.64
male (%)	50.10	1.89	45.02	67.60
less than age 55 (%)	69.61	6.17	45.07	86.12
white (%)	84.88	16.39	10.73	99.45
avg. household size	2.50	0.25	1.95	4.05
with bachelor's degree or higher (%)	20.01	7.80	6.23	64.01
unemployed (%)	7.00	3.69	0.61	28.98
median household income (US-$)	46,296.10	10,436.37	21,399	105,989
Gini index of inequality	0.44	0.03	0.34	0.56
owner-occupied housing units (%)	71.85	7.44	40.70	89.99
median rent as proportion of income (%)	27.43	4.43	10.00	44.70
avg. commute time to work (minutes)	21.28	5.30	10	42

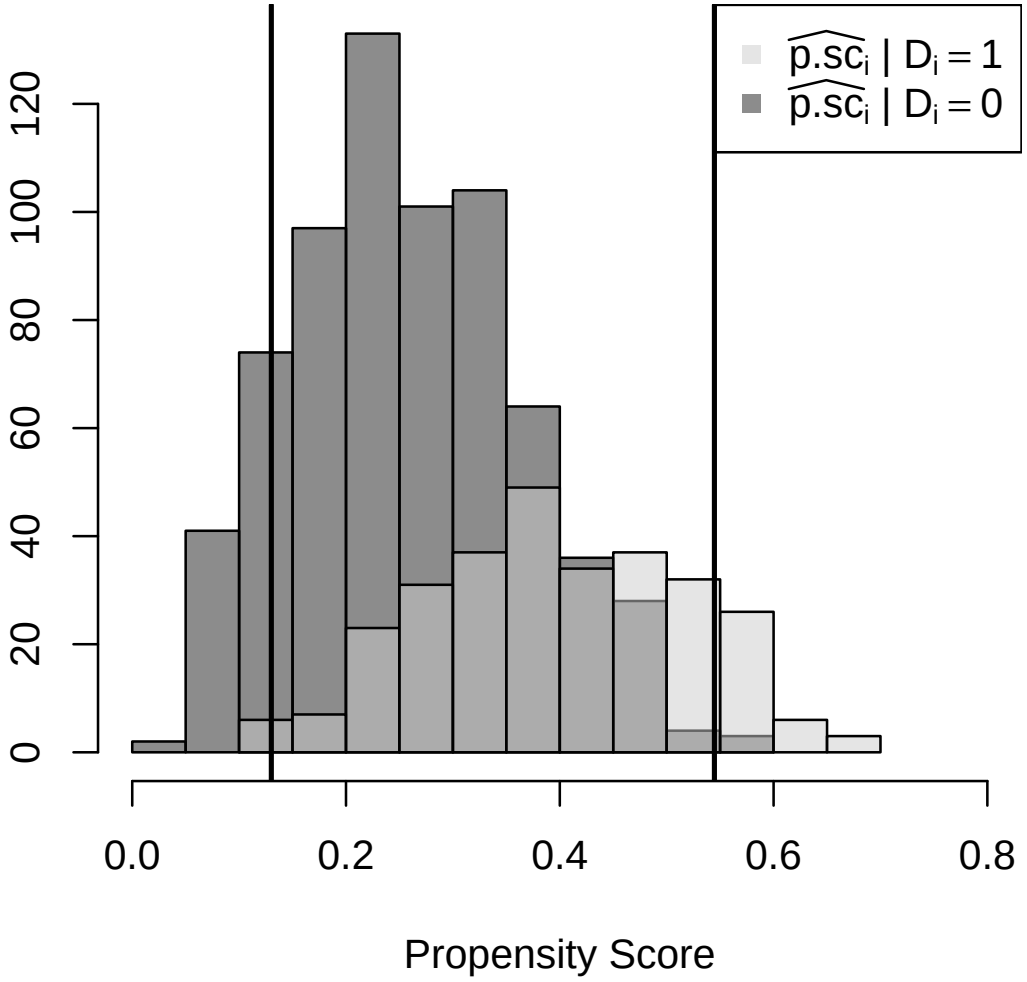

Figure C.1: Distribution of propensity score estimates by treatment status. Regions of overlap and non-overlap are indicated by the vertical straight lines.

D SBART+SPL for binary outcomes

The SBART+SPL algorithm described in Algorithm 2 can be adapted to handle binary instead of continuous outcomes by implementing two trivial changes following Nethery et al. (2019). First, the treatment effects in Definition 2.1 have to be redefined in the presence of binary outcomes as in Definition D.1. Second, a SoftBART probit link function is used in the region of overlap to estimate individual causal effects to adapat for the case that binary outcomes can just take on values of -1, 0, 1. Third, the estimates are bounded between -1 and 1 using arcsine transformation on the SoftBART probit estimates.

Definition D.1. Redefinition of treatment effects for binary outcomes

$$\tau_i = \Pr\left(Y_i(1) = 1\right) - \Pr\left(Y_i(0) = 1\right), \tag{D.1}$$

$$\tau_S = \frac{1}{n}\sum_{i=1}^{n} \tau_i, \tag{D.2}$$

$$\tau_{P|\mathbf{x}} = \Pr\left(Y(1) = 1|\mathbf{X} = \mathbf{x}\right) - \Pr\left(Y(0) = 1|\mathbf{X} = \mathbf{x}\right), \tag{D.3}$$

$$\tau_P = \mathbb{E}_{\mathbf{X}}[\tau_{P|\mathbf{x}}]. \tag{D.4}$$

More precisely, the model in Equation (4.3) for imputation of missing outcomes reads for binary outcomes, with $\Phi(.)$ being the standard Normal cumulative distribution function, as

$$Pr\left(Y_o^{obs} = 1\right) = \Phi\left(\sum_{j}^{J} g\left(D_o, \widehat{p.sc}_o, \mathbf{X}_o; \mathcal{T}_j, \mathcal{M}_j\right) + \epsilon_o, \epsilon_o \sim \mathcal{N}(0, \sigma^2_{Imp})\right), \tag{D.5}$$

and a SoftBART probit model is used to retrieve posterior samples $\tilde{Pr}\left(Y_o^{obs}\right), \tilde{Pr}\left(Y_o^{mis}\right)$ from the corresponding posterior predictive distributions to compute $\tilde{\mathbf{\Delta}}_O$.

In the extrapolation phase, an arcsine transformation is applied to the individual causal effects in the region of overlap such that the smoothing model in Equations (4.6) and 4.7 change to

$$\arcsin\left(\tilde{\Delta}_o\right) = \mathbf{W}'\boldsymbol{\beta}_{Smo} + \epsilon_o \ , \ \epsilon_o \sim \mathcal{N}\left(0, \sigma^2_{Smo}\right) \tag{D.6}$$

with

$$\mathbf{W} = \begin{cases} \left[rcs\left(\widehat{p.sc}_o\right) \ , \ Y_o^\star(1) \ , \ \mathbf{X}_o\right]', \ \text{if } D_o = 1 \\ \left[rcs\left(\widehat{p.sc}_o\right) \ , \ Y_o^\star(0) \ , \ \mathbf{X}_o\right]', \ \text{if } D_o = 0 \end{cases} . \tag{D.7}$$

Moreover, the Equations (4.8) and (4.9) are replaced as follows: For treated individuals in the non-overlap region ($D_{o^\neg} = 1$), one uses

$$Y_o^\star(1) = \begin{cases} Y_o^{obs} \ , \text{ if } D_o = 1, \\ \mathbb{1}\left(\tilde{P}\left(\tilde{Y}_o^{mis} = 1\right) > 0.5\right) \ , \text{ if } D_o = 0 \end{cases} . \tag{D.8}$$

For non-treated individuals in the non-overlap region ($D_{o^\neg} = 0$), one uses

$$Y_o^\star(0) = \begin{cases} Y_o^{obs} \ , \text{ if } D_o = 0, \\ \mathbb{1}\left(\tilde{P}\left(\tilde{Y}_o^{mis} = 0\right) > 0.5\right) \ , \text{ if } D_o = 1 \end{cases} . \tag{D.9}$$

In the end, the extrapolated individual causal effects in region of non-overlap that are drawn from the posterior predictive distribution, as described in Equation (4.10), have to be back-transformed from the arcsine scale by applying $\sin\left(\tilde{\Delta}_{o^\neg}\right)$.